INTEGRATED CORRELATION ANALYSIS OF PROTEOMICS AND TRANSCRIPTOMICS DATA IN ALZHEIMER'S DISEASE

Suneeta Modekurty

MS Bioinformatics

Submitted to the faculty of the University Graduate School in partial fulfillment of requirements for the degree Master of Science

in Bioinformatics in the School of Informatics and Computing,

Indiana University, IUPUI

December 2020

Accepted by the Graduate faculty, Indiana University, in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics

Master's Thesis Committee

Xiaowen Liu

Xiaowen Liu, Ph.D.

Jun Wan

Jun Wan, Ph.D.

Jiaping Zheng, Ph.D.

Acknowledgment

Acknowledgment

Firstly, I would like to thank my thesis advisor, Dr. Xiaowen Liu, for setting specific aims and guiding me. Without his direction, encouragement and constructive criticism, this thesis would not have been possible. My sincere thanks to our committee members, Dr Jun Wan and Dr Jiaping Zheng for agreeing to be part of my defense and giving helpful suggestions. I would like to thank the Department of Bio-Health Informatics, School of Informatics and Computing at Indiana University-Purdue University at Indianapolis for giving me the opportunity to pursue a Master of Science in Bioinformatics. Lastly, I would extend my heartfelt thanks to my husband, children and friends for constantly supporting and loving me every step of the way.

INTEGRATED CORRELATION ANALYSIS OF PROTEOMICS AND TRANSCRIPTOMICS DATA IN ALZHEIMER'S DISEASE

Abstract

Neurodegeneration is the umbrella term for a range of conditions which include various types of conditions like Progressive Supranuclear Palsy (PSP), Alzheimer's disease (AD), and Huntington's disease which involve the death of neurons. Once the neurons die, they cannot be replaced. The Alzheimer's Association has mentioned in their facts and figures that Alzheimer's disease (AD) is thought to begin at least 20 years before the symptoms are seen. The Association further categorized AD into five broad stages based on the CERAD score [1]. AD destroys cognitive functioning and thinking as it progresses. However, disease-modifying treatment is yet not there [2,3,4]. For some years now, researchers have successfully worked on the disease with individual omics layers. With the rise in next-generation sequencing technologies and the reduced cost of sequencing, it is now possible to integrate large datasets that facilitate information pertaining to biology. Thus, it is now a known fact that there is an excellent interplay between the omics layers contrary to what is laid down by the central dogma [5]. However, co-immunoprecipitation methods can confirm the actual cross talks. There are also numerous integrated studies like network analysis, principal component analysis, clustering, correlation analysis, and predictive analysis to understand the associations between the different omics layers. We wanted to see if there existed any significant correlations between two -omics layers. So, here, we performed a correlation analysis to study the disease. The pipeline building consisted of first

performing the differential expression of two datasets (proteomics and transcriptomics) individually. An in-depth analysis of the proteomics data was performed, followed by differential expression analysis of RNA seq data and then a correlational analysis of the differentially expressed proteins (from proteomics data) and genes (from RNA seq data). From our analysis, we found fascinating information about the correlations between proteins and genes in AD. We performed a correlation analysis of AD (N= 84), Control (N = 31), and PSP (N = 85) samples for proteomics data and got 114 differentially expressed proteins (DEPs = 114). The RNA seq data had AD (N = 82), Control (N = 31) and PSP (N = 84) samples which gave us 61 differentially expressed genes (DEGs = 61). A correlation analysis using Spearman's correlation coefficient method between proteins involved in AD revealed 192 very significant correlations with p-value <= 0.00000000000005. The mean correlation coefficient was quite high (r = 0.52). A correlation analysis using Spearman's correlation coefficient method between genes involved in AD revealed 208 very significant correlations with p-value <= 0.00000000000005. The mean correlation coefficient was quite high (r = 0.52). A correlation analysis using Spearman's correlation coefficient method between proteins and genes involved in AD revealed 395 significant correlations with p-value <= 0.0001. The correlation coefficient (quite high of +0.53), which might help in understanding the molecular pathways behind the disease could uncover new prospects of understanding the disease as well as design treatments. We observed that different genes interact with different proteins (correlation coefficient r >= 0.5, p-value < 0.05). We also observed that a single protein interacts with multiple genes, and a single gene is interestingly associated with multiple proteins. The patterns of correlations are also different in that a protein/gene positively correlates with some proteins/genes and negatively with some other proteins/genes. We hope that this observation is quite useful. However, understanding how it works and how they interact with each other needs further assessment at the molecular level.

Contents

List of Figures

Chapter 1

Chapter 2

List of tables

Chapter 2

Chapter 1: Introduction

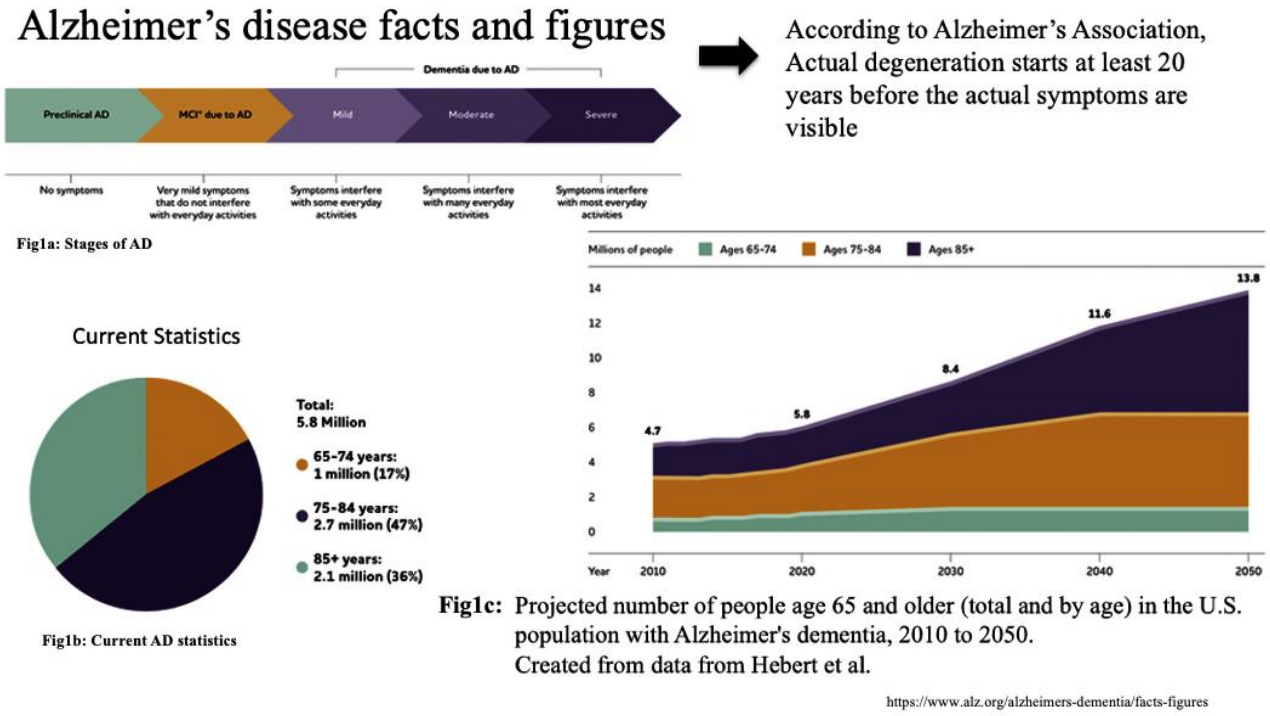


Fig1: AD facts and figures as given by Alzheimer's Association

The data from statistical facts and figures for 2020 by Alzheimer's Association shows that at least 5.8 million Americans have Alzheimer's Disease (AD), of which 1 million (17%) are in the age range of 65-74 years, 2.7 million (47%) are in between 75-84 years and 2.1 million (36%) are above 85 years of age. This number is thought to reach 13.8 million by 2050. Alzheimer's Association has mentioned in its facts that AD is thought to begin at least 20 years before the symptoms are seen. The Association further categorized AD into five broad stages based on the CERAD score [1]. Stage-1 is the Preclinical AD with no symptoms, stage-2 is a mild cognitive impairment (MCI) that do not interfere with daily activities, stage-3 is mild but interferes with daily activities, stage-4 is moderate AD interfering with most daily activities, and the final stage is stage-5 in which the symptoms interfere with all daily activities. As the disease progresses, the structure of the brain changes. Currently, there are

symptomatic treatments available for AD, namely donepezil, rivastigmine, galantamine, and memantine, but they do not play a role in actually decelerating or preventing the progression of the disease [3]. While protein aggregation is the hallmark of any form of neurodegeneration (Lee et al., 2011), the disease's molecular epidemiology is still under debate. Researchers believe that the disease begins when amyloid-beta begins to accumulate. Beta-amyloid 42, formed from the proteolytic breakdown of amyloid precursor protein (APP), is a very toxic form of beta-amyloid, as stated by the National Institute of Aging. Abnormal quantities of this naturally occurring protein form plaques between the synapses, causing a neurodegenerative condition leading to AD. Abnormal accumulation of neurofibrillary tangles called Tau accumulation inside the neurons due to the hyperphosphorylation of microtubule-associated protein (MAPT) is also a cause of AD. The beta-amyloid is normally cleared off by the microglia, but sometimes, too much is released, and among the synapses and not enough of it is cleared away [4,5]. Like AD, PSP also involves the abnormal deposition of hyperphosphorylated tau. The difference is that abnormal deposition of amyloid-beta plaques is absent in PSP. PSP and AD also differ in the type of tau protein deposition. The AD-type hyperphosphorylated Tau is different from PSP-type hyperphosphorylated Tau in its isomers formed due to alternative splicing [6]. There are numerous integrated studies done to understand the interactions occurring between the different omics layers. For example, (Singhal et al., 2019) proposed a pathway-centric neural-network-based omics integration analysis [6]. (Dragomir et al., 2019) reviewed the current state-of-the-art integrated framework focusing on finding network-based biomarkers at molecular and brain functional connectivity levels [7]. (Avramouli & Vlamos, 2016) highlighted the most recent next-generation sequencing technologies that have enabled a sophisticated analysis of the human genome [8]. A correlation analysis is a very good way to look at how two quantitative variables form pairs with each other. Here, we aim to perform a correlation analysis of proteomics and transcriptomics data for AD

and PSP. Mathematically, correlation is defined as the degree of relationship or association between two variables. When two variables or two datasets show a high correlation, we can infer that they are closely linked. So, when we do a correlation, we can either look at how one variable tends to change with respect to the other, or how both the variables change together. Sometimes If one variable is large, another variable may become smaller or larger. Also, the relationship between two variables can be positive if high measures of one variable correspond to high measures of another variable or the relationship between two variables can be negative if high measures of one variable correspond to low measures of another variable (Gonick et al., 2003). In any case a correlation may not be causation. So, there can exist a high correlation between two variables with no causations at all. They are just the predictors of each other. The + or - of the correlation coefficient (which depicts both a direction and magnitude) tells that there exists a relationship between the two, with + being a positive relationship, - being a negative relationship, and the number indicating the magnitude or the value of the relationship. Many studies show that there is a very poor correlation between mRNA and proteins. The discrepancy is typically attributed to regulation levels between transcript and protein products (Maier et al., 2009). Advances in next-generation sequencing technologies and methods and mass spectrometry proteomics provide an unparalleled ability to survey mRNA and protein abundances helping researchers to explore the extent to which different aspects of gene expression help to regulate cellular protein abundances (Vogel et al., 2012). Data demonstrates that the regulatory processes that occur after mRNA are made are myriad, like post-transcriptional regulation, post-translational regulation, and protein degradation regulation (Fekete et al., 2012), controlling steady-state protein abundances. Most mRNA-protein correspondence studies calculate a single correlation coefficient representing a correlation between mRNA expression and protein expression across all genes. The correlation between mRNA and protein experiments may represent a general measure of how well mRNA and protein expression corresponds

across the entire genome. The correlation coefficient represents the correlation between the expression of an mRNA and its protein product across multiple samples or conditions (Koussounadis et al., 2015). So, since our focus was to find correlations between proteins (proteomics) and genes (transcriptomics) and if there are any patterns to their pairing, we performed a correlation analysis using Spearman's coefficient method.

Chapter 2: Materials and Methods

2.1: Experimental design and data download:

The data sets were downloaded from open-source platform synapse.org-MayoRNAseq study (Synapse ID: syn5550404). This study is independent of studies described under the Mayo Clinic Alzheimer's Disease Genetics Studies (MCADGS). The data in the MayoRNAseq study consists of whole transcriptome data for 276 Temporal cortex (TCX) samples from 312 North American Caucasian subjects. The brain samples are a single cohort diagnosed with neuropathological conditions of Alzheimer's Disease (AD) and Progressive Supranuclear Palsy (PSP). Controls (C) were those elderly patients that did not have any neurodegenerative condition. Within this cohort, all AD, PSP and Control subjects were from the Mayo Clinic Brain Bank (MCBB). All subjects selected from the MCBB underwent neuropathologic evaluation [10]. All ADs had definite diagnosis according to NINCDS-ADRDA criteria and had a Braak NFT stage of IV or greater [11]. Control subjects had Braak NFT stage of III or less, CERAD neuritic and cortical plaque densities of 0 (none) or 1 (sparse) and lacked any of the neurodegenerative pathologic diagnoses including AD, Parkinson's disease (PD), PSP, motor neuron disease (MND), Pick's disease (PiD), Huntington's disease (HD), hippocampal sclerosis (HipScl), dementia lacking distinctive histology (DLDH) or any other forms of dementias [10]. Under mentioned (**Table 1**) is the sample sizes for proteomics and transcriptomics data and their description used in this project.

Table 1: Summary of the datasets used in the analysis:

Origin of samples	Dataset	Conditions	No. of Samples
Postmortem adult brains	**Proteomics**	**AD** **PSP** **Control**	**84** **85** **31**
Postmortem adult brains	**Transcriptomics**	**AD** **PSP** **Control**	**82** **84** **31**

2.1.1: Proteomics dataset description:

The brain samples of elderly adults whose age was in the range of 65-90, Alzheimer's disease (AD; N=84), progressive supranuclear palsy (PSP; N=85), pathologic aging control (CON; N=31), were quantified using label-free (LFQ) based protocols followed by LC MS/MS analysis [10]. The output proteomics abundance dataset, "Mayo_proteomics_TC_proteinoutput.txt" and associated traits, "Mayo_proteomics_TC_traits.csv" were used for further bioinformatics and statistical analysis. RAW data for the samples was analyzed using MaxQuant v1.5.3.30 with Thermo Foundation 2.0 for RAW file reading capability. The search engine Andromeda, integrated into MaxQuant 1, was used to build and search a concatenated target-decoy Uniprot human reference protein database (retrieved April 20, 2015; 90,303 target sequences), plus 245 contaminant proteins from the common repository of

adventitious proteins (cRAP) built into MaxQuant. Methionine oxidation (+15.9949 Da), asparagine and glutamine deamidation (+0.9840 Da), and protein N-terminal acetylation (+42.0106 Da) were variable modifications (up to 5 allowed per peptide); cysteine was assigned a fixed carbamidomethyl modification (+57.0215 Da). Only fully tryptic peptides were considered with up to 2 mis-cleavages in the database search. A precursor mass tolerance of ±20 ppm was applied prior to mass accuracy calibration and ±4.5 ppm after internal MaxQuant calibration. Other search settings included a maximum peptide mass of 6,000 Da, a minimum peptide length of 6 residues, 0.05 Da tolerance for high resolution MS/MS scans. Co-fragmented peptide search was enabled to deconvolute multiplex spectra. The false discovery rate (FDR) for peptide spectral matches, proteins, and site decoy fraction were all set to 1 percent. Quantification settings were as follows: re-quantify with a second peak finding attempt after protein identification has completed; match MS1 peaks between runs; a 0.7 min retention time match window was used after an alignment function was found with a 20-minute RT search space. Quantitation of proteins was performed using summed peptide intensities given by MaxQuant. The quantitation method only considered razor plus unique peptides for protein level quantitation. The full list of parameters used for MaxQuant are available as Mayo_Proteomics_TC_searchparameters.xml accompanying the public release [10]. The data has Synapse ID: syn7431760.

2.1.2: Transcriptomics dataset description:

Transcript read count data was available to be downloaded (Synapse ID: syn20818651). Gene expression measures were generated using next-generation RNA sequencing (RNAseq), at the MCBB Sun Health research institute from temporal cortex(TCX) for 278 subjects. Out of these 278 subjects, 82 subjects were diagnosed with Alzheimer's disease (AD), 84 were diagnosed with progressive

supranuclear palsy (PSP), 28 were diagnosed with pathologic aging (PA), and 80 were control (C) samples.

Bioinformatics methods summary: SNAPR aligner was used for read alignment. The human reference genome version used for indexing were GRCh38 reference and Ensembl v77 gene models. The output files were gene counts and transcript counts for each sample. We considered the transcript counts. QC was done to identify any subjects with mapped reads less than 85% or any sex-dependent (Y chromosome gene expression) using PLINK [13]. To identify any population outliers (defined as > 6 standard deviations from the mean after 5 iterations), EIGENSTRAT was used [14]. All reads have PHRED score >= 20 [13,14,15,16].

To identify the genes that are differentially expressed across different conditions, a differential expression analysis using R package edgeR(empirical analysis of DGE in R) was performed. EdgeR also works like limma but mathematically is more complex (Robinson et al., 2010). Those differentially expressed genes were used for correlation analysis. PA (N = 84) samples were removed from the dataset to match the proteomics data set which had AD, PSP and C samples and lacked PA samples. For controls, 31 samples that matched the proteomics dataset were considered for differential expression analysis. Thus, we now had Alzheimer's disease (AD; N=82), progressive supranuclear palsy (PSP; N=84), and control (C; N=31) samples in our transcriptomics data set.

2.1.3: Workflow employed in the analysis of the project:

Following the fetching of data for proteomics and transcriptomics from synapse.org, we designed the workflow depicted below. In the first step, we cleaned and manipulated proteomics data (removed contaminants) performed batch corrections followed by a differential expression analysis, inferential

statistical analysis and protein co-expression analysis. RNA seq data was also processed in parallel and differential expression analysis was done. A correlation analysis of the differentially expressed proteins and genes was then carried out.

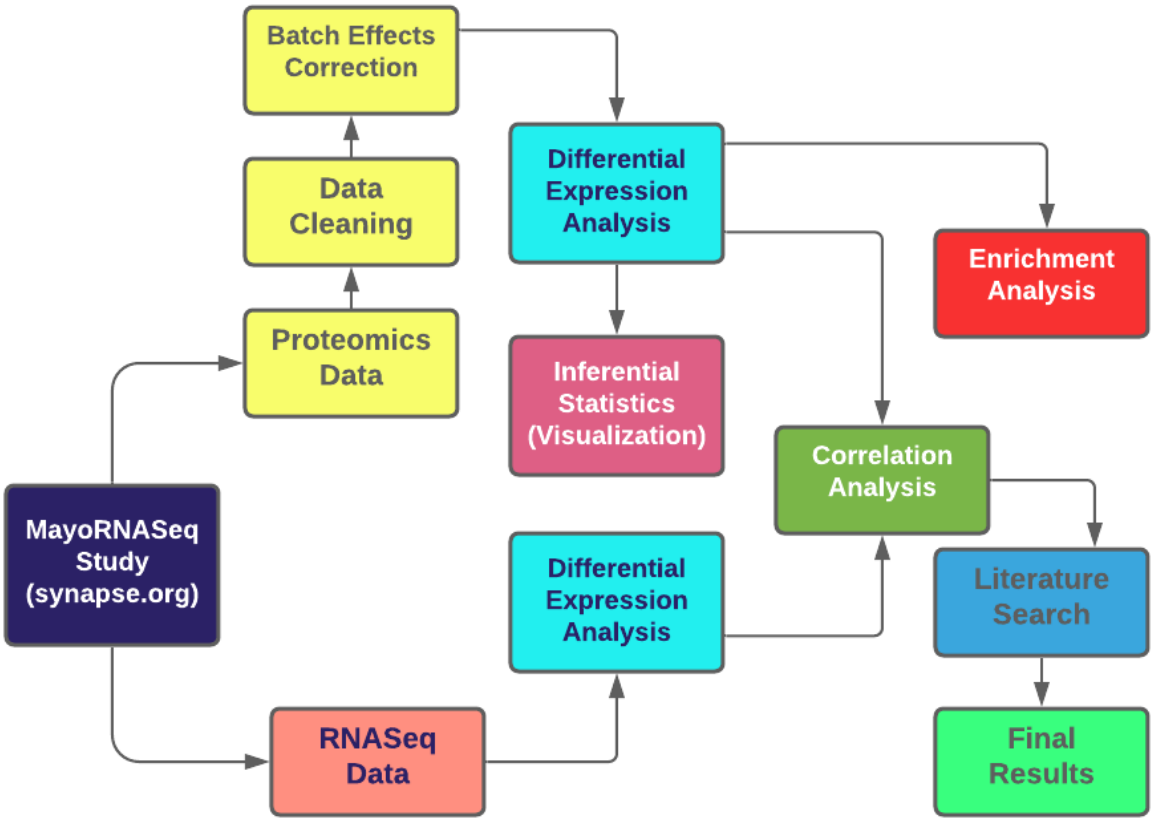


Figure 2: Workflow employed to identify proteins and genes correlated in our analysis.

2.2: Proteomics data analysis:

2.2.1: Background:

In a review paper (Lee et al., 2018) mentioned that neurodegenerative diseases like AD are not only painful to the person suffering from it but also a burden to the patient's social relationships. They further assert that mass spectrometry is a robust technology to study protein abundances involved in a disease condition in which proteins are digested by trypsin and then analyzed by LC-MS/MS. A lot of research from the past (www.mayoclinic.org) suggests that a person who is the carrier e4 variant of the APOE gene (APOE e4) is at the highest risk of AD. Microtubule-associated protein tau (MAPT) is the major protein of a mature neuron, others being MAPT1 and MAPT2, which appear to be critical to regular brain activity in evolved species like humans. MAPT has associated protein variants like MAP1A, MAP1B and MAP2. So, if microtubule associated protein tau (https://www.genenames.org/tools/search/#!/all?query=MAPT) is non-functional, its functionality can be compensated by the other associated proteins. However, it is the toxicity of tau protein that affects a neuron's functionality, leading to a gradual eroding of the brain's structure leading to eventual functional loss. The hyperphosphorylation of tau protein is a result of an imbalance of activities between kinases and phosphates [16]. It is now clear that protein aggregation is the hallmark of neurodegeneration [17]. Here, to identify and to understand the molecular associations involved in AD, we performed an in-depth proteomic analysis of postmortem human elderly brains and later a correlation analysis of the proteins to understand the pairing of a protein with its neighboring proteins in AD with respect to PSP, controls and also with respect to each other (AD). We performed a pairwise comparison of AD vs C, PSP vs C and AD vs PSP when performing a differential expression analysis.

Table 2: Summary of samples used in proteomics analysis:

Data	Batches	Conditions	Rows/Number	Columns/Number
Proteomics	b1	AD	Human proteins (6585)	Samples - Label-Free Quantification With intensities for the three conditions (2105)
	b2	PSP		
	b3			
	b4	Control		
	b5			

2.2.2: Data cleaning and preprocessing - Batch effects correction:

Batch effects in proteomics often result in both, decreased intensities and increased number of missing values [17]. The protein abundance data which is an output data of MaxQuant is a single cohort with 5 batches (namely b1, b2, b3, b4, b5) and 3 conditions (AD, Control, PSP). The dataset has 6585 rows representing human proteins and 2105 columns representing Label-free quantified (LFQ) intensities and other related information which included contaminants, m/z values etc. After filtering contaminants, there are 6335 human proteins in the rows and 2105 columns which included LFQ intensities along with other important information pertaining to the proteins. Within each batch, the data had pooled samples named after “mgis” that were the Mayo Global Internal Standards and “egis” that were the Emory Global Internal Standards as can be seen in table 1. Pooled samples are often added to proteomics experiments in order to overcome resource constraints when many individuals are analyzed and to reduce biological variance [18,19].

For example, if σ_b^2 is biological variance and σ_t^2 is technical variance, then the expected variance in expression of protein in a sample of individuals, σ_i^2 is,

$$\sigma_i^2 = \sigma_b^2 + \sigma_t^2 \qquad (1)$$

The variance in a sample of pools σ_p^2 each formed by combining equal amounts of total protein from r individual samples is,

$$\sigma_p^2 = \frac{1}{r}\sigma_b^2 + \sigma_t^2 \qquad (2)$$

Thus Eq. (2) shows that the measured biological variance in a pool will decrease by a fraction 1/*r*. This reduction should increase the power to detect treatment differences.

Pooled samples are normally prepared by taking a little (equal) amount of each sample, which is put into two aliquots and digested along with other samples, to serve as pooled Global Internal Standards (GIS) [20]. Though pooled samples reduce the biological variance, they are a major source of artifacts and batch effects [21]

Table 3: Pooled samples with a suffix of “egis” and “mgis” in five batches (b1, b2, b3, b4, b5)

Samples_Simple	samples	batch	batch
b1_02	b1_02_egis	b1	EmoryGlobalInternalStandard
b1_24	b1_24_egis	b1	EmoryGlobalInternalStandard
b1_46	b1_46_egis	b1	EmoryGlobalInternalStandard
b1_01	b1_01_mgis	b1	MayoGlobalInternalStandard
b1_23	b1_23_mgis	b1	MayoGlobalInternalStandard
b1_45	b1_45_mgis	b1	MayoGlobalInternalStandard
b2_02	b2_02_egis	b2	EmoryGlobalInternalStandard
b2_24	b2_24_egis	b2	EmoryGlobalInternalStandard
b2_46	b2_46_egis	b2	EmoryGlobalInternalStandard
b2_01	b2_01_mgis	b2	MayoGlobalInternalStandard
b2_23	b2_23_mgis	b2	MayoGlobalInternalStandard
b2_45	b2_45_mgis	b2	MayoGlobalInternalStandard
b3_02	b3_02_egis	b3	EmoryGlobalInternalStandard
b3_24	b3_24_egis	b3	EmoryGlobalInternalStandard
b3_46	b3_46_egis	b3	EmoryGlobalInternalStandard
b3_01	b3_01_mgis	b3	MayoGlobalInternalStandard
b3_23	b3_23_mgis	b3	MayoGlobalInternalStandard
b3_45	b3_45_mgis	b3	MayoGlobalInternalStandard
b4_02	b4_02_egis	b4	EmoryGlobalInternalStandard
b4_24	b4_24_egis	b4	EmoryGlobalInternalStandard
b4_46	b4_46_egis	b4	EmoryGlobalInternalStandard
b4_01	b4_01_mgis	b4	MayoGlobalInternalStandard
b4_23	b4_23_mgis	b4	MayoGlobalInternalStandard
b4_45	b4_45_mgis	b4	MayoGlobalInternalStandard
b5_02	b5_02_egis	b5	EmoryGlobalInternalStandard
b5_24	b5_24_egis	b5	EmoryGlobalInternalStandard
b5_45	b5_45_egis	b5	EmoryGlobalInternalStandard
b5_01	b5_01_mgis	b5	MayoGlobalInternalStandard
b5_23	b5_23_mgis	b5	MayoGlobalInternalStandard
b5_44	b5_44_mgis	b5	MayoGlobalInternalStandard

2.2.3: Controlling for batch-specific variance:

There were severe batch effects across the five batches in our proteomics dataset and the data needed to be cleaned. The data was pre-processed and cleaned by using TAMPOR MASTER which normalized the data and fixed the batch effects. (Tunable Approach for Median Polish of Ratio -- Biological Abundance Batch Effect Removal). (https://github.com/edammer/TAMPOR) [22]. TAMPOR (available as an R script) implements a median polish algorithm to adjust technical variance due to multiple samples, cohorts, or batches. The algorithm implements several iterations of equation (1) below.

$$\frac{abundance}{median(AllSamples)batch} \times \frac{grand\ median}{median\left(\left\{\frac{abundance}{median(AllSamples)batch}\ \text{All samples from batch}\right\}\right)} \quad (1)$$

The algorithm implements equation (1) for each protein abundance measurement across each sample for all samples converting them into median-centered abundance measurements. The matrix is then log2 transformed. Each log2 ratio is then adjusted by subtracting the sample log2 ratio median for each protein. The ratios are then anti-logged. Row wise median of all samples is then multiplied to the anti-logged ratios. This process is iterated until convergence. The algorithm reduces technical variance and preserves the biological variance [23]. It works well even if batch replicates for normalization have unusual variance compared to biological samples not used for normalization (e.g., due to differential peptide digestion, different tissue region, or different genetic background of control samples) [23]. The output obtained was a normalized dataset which was used to carry out the rest of the analysis. As mentioned in Table 1, there were two GIS samples ("mgis" and "egis"). We considered "mgis" samples for removal of batch effects. TAMPOR also gave the plots before and after batch effect removal. In figure 3, as we can see, the left plot is the plot of raw data with a lot of batch effects and is not clean. It

is very discrete and heterogenous. But on the right is the plot after batch effects were removed. Now the data was ready for further analysis.

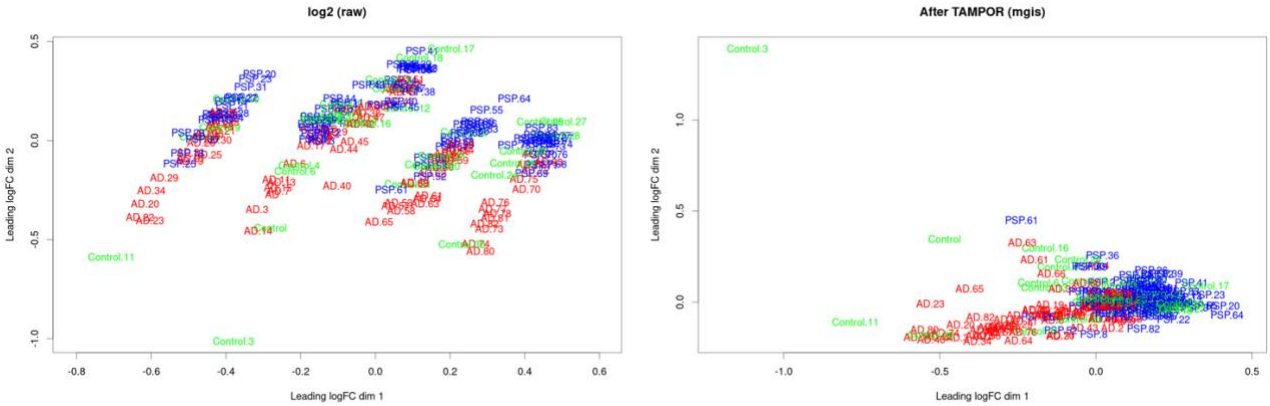


Figure 3: MDS plot of the data before and after batch effect removal. Colour scheme: Green = Control (31), Red = AD (84), Blue = PSP (85).

Since the proteomics dataset is prepared using label free quantification methods, we needed only LFQ intensity values for all five batches from the data set. These LFQ intensity values were selected and extracted for AD, PSP and C samples. So, now our data had 6335 rows representing proteins and 215 columns with LFQ intensities (of samples). The Traits file has lines 1:200 – samples (LFQ intensities), lines 201:215 – egis samples (removed), lines 216:230 – 'mgis' samples.

2.2.4: Missing values and imputation:

The abundance of missing values due to technical or biological reasons is a prominent characteristic of proteomics data that can bias every normalization procedure and violate the distributional assumptions that are foundations of normalization [24,25]. So, before any normalization is carried out, missing values must be tackled. The missing values in the dataset were removed while correcting batch effects in which we eliminated all rows having more than 50% missing values. The data still had some missing

values and was imputed by replacing the missing values with the median observation for each protein under each condition.

2.2.5: Quantile normalization:

The aim here is to give different distributions the same statistical properties by removing the technical variations between experiments. Mass spectroscopy-based quantification often compromises on the normalization steps before the downstream analysis is done. Normalizing proteomics data is necessary to remove the biases that arise due to non-biological variance/technical artifacts. [24,25]. Quantile normalization (QN) method was used to remove such unintended variation while retaining the biological signal of interest. QN considers the mean of the most expressed values across all samples and makes an empirical distribution of abundance values for each sample to be the same. So, the normalized dataset retains the rank of the protein/gene expression values as well as removes variance among the expression values. QN was applied to a matrix of proteins as rows and samples with LFQ-intensity values as columns. The steps included sorting the proteins for each sample, finding the rank of values in each column, finding row mean on sorted columns, substituting mean values according to rank, and replacing each column's value with mean according to rank. Here, by rank, we mean sorting the protein abundances according to their expression values, highest being on the top.

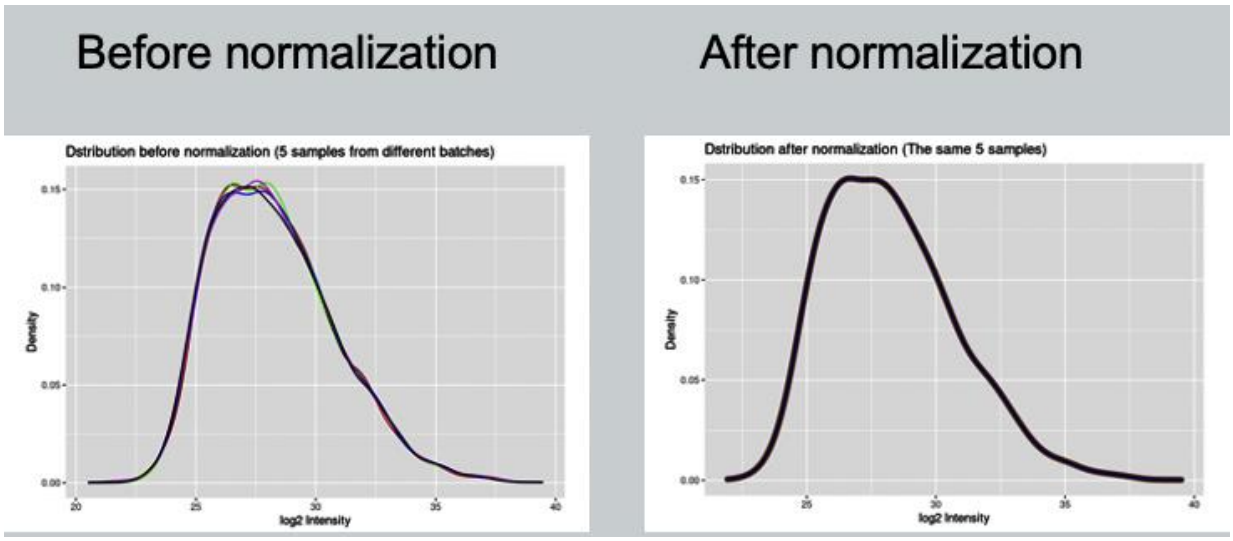


Figure 4: Plots before and after quantile normalization of the data

2.2.6: Differential expression analysis:

To analyze mass-spectrometry based proteomics data in which an effort is being made to compare disease to normal samples, a standard statistical method must be employed to obtain significant differentially expressed proteins (DEPs) across different experiments (https://www.sciencedirect.com/science/article/pii/S2212968515000069#bib0065).

We used the Bioconductor package 'limma' to perform proteomics data analysis. Limma uses the Empirical Bayes method to analyze data (Smyth, G. K. (2004). Limma makes a pooled estimate to retain biological variance in the data by reducing the sample variances. A pooled estimate is an estimate obtained when combined information from two or more randomly selected samples having the same mean is taken. Thus, the analysis becomes fully unbiased and more powerful than normal t-tests. Analysis with limma follows the under mentioned steps.

First, the pooled variance is computed by computing within each protein's region variance and averaged across all proteins. The pooled variance is computed by creating a design matrix (A design matrix is a matrix whose columns give the coefficients of the linear model. There is one row for each sample) from traits data, computing the within-sample correlation for each protein and fitting the model to obtain the pooled variance. Next, a t-test is computed for each protein to get a list of significant DEPs based on p-values, adjusted p-values, or FDR estimates. We already had the dataset that was corrected for batch effects. Before performing the DEP analysis, we further imputed and normalized (QN) the dataset which was the input for further DEP analysis. As was essential, a design matrix was created, and linear model fitting was done. We performed a pairwise comparison of the experimental conditions with AD vs. Control, PSP vs. Control and AD vs. PSP. So, we had to build a contrast of each pairwise sample. Contrasts.fit() function was used to compute estimated coefficients and standard errors for a given set of contrasts (here, AD vs Control). The input for the contrasts.fit() was lmFit.

2.2.6.1: Statistics for Differential Expression:

Limma employs eBayes() function which returns a number of summary statistics for each protein. The logFC (log2-fold change) is the log2-expression level for that protein. AveExpr, the average expression level for that protein across all the samples. t is t-statistic which is the ratio of the logFC to its standard error. (t-statistic is the same as an ordinary t-statistic but the standard errors are moderated across proteins, so that information about each protein is an ensemble of information from all proteins. The p-value is then obtained from t-statistic. Since the number of samples is many, multiple hypothesis correction is necessary to reduce the false positives. The FDR adjustment method used here was 'BH' (Benjamini and Hochberg's method). The B-statistic talks about the log-odds that a particular protein is differentially expressed and is automatically set to 1% for multiple testing. Of all the statistics,

considering p-value (adj-p-value) to select differentially expressed proteins is a general practice. Limma's eBayes() thus returned 405 proteins significantly downregulated and 321 upregulated in AD vs. C samples. 346 proteins were significantly downregulated and 521 upregulated in PSP vs. C samples and 1240 proteins are significantly downregulated and 823 upregulated in AD vs. PSP samples as depicted in table 3.1, table 3.2 and table 3.3 respectively.

Table 4.1: Significant DEPs in the comparison of proteins in AD vs Control samples

PID	logFC	AveExpr	t	P.Value	Adj.P.Val	B
H0YG30	0.685	27.101	8.127	0.00E+00	0.00E+00	21.349
A6NMN0	0.833	29.329	7.28	0.00E+00	0.00E+00	16.488
VAMP3	0.463	25.338	6.919	0.00E+00	0.00E+00	14.501
TAU.1	0.844	30.912	6.673	0.00E+00	0.00E+00	13.188
PLEC	0.537	25.819	6.567	0.00E+00	0.00E+00	12.63
H0YFX9	0.462	30.405	6.542	0.00E+00	0.00E+00	12.498
H7C545	0.829	29.007	6.306	0.00E+00	0.00E+00	11.278
H0Y7G9	0.809	26.566	6.075	0.00E+00	0.00E+00	10.117
X6RHB9	0.403	27.84	6.056	0.00E+00	0.00E+00	10.023
E9PQN5	-0.384	25.91	-5.923	0.00E+00	0.00E+00	9.366
AB17A	0.631	23.663	5.806	0.00E+00	0.00E+00	8.8
H0YJI1	-0.322	27.909	-5.805	0.00E+00	0.00E+00	8.794
A0A087X134	0.356	32.848	5.788	0.00E+00	0.00E+00	8.709
E5RG36CON__P17697	0.367	31.652	5.746	0.00E+00	0.00E+00	8.511
CO4A	0.844	29.153	5.636	0.00E+00	0.00E+00	7.987

Table 4.2: Significant DEPs in the comparison of proteins in PSP vs Control samples

PID	logFC	AveExpr	t	P.Value	adj.P.Val	B
H0YHD8	0.373	31.417	6.349	0.00E+00	0.00E+00	11.473
E9PR42	-0.461	27.668	-6.107	0.00E+00	0.00E+00	10.258
E9PLK6	-0.459	30.495	-6.06	0.00E+00	0.00E+00	10.022
ANO1	0.509	28.938	5.873	0.00E+00	0.00E+00	9.112
X6RHB9	0.383	27.84	5.767	0.00E+00	0.00E+00	8.601
J3QRJ3	0.305	33.287	5.745	0.00E+00	0.00E+00	8.498
ARHG6	-0.302	25.122	-5.735	0.00E+00	0.00E+00	8.45
ARAID	-0.556	37.127	-5.644	0.00E+00	0.00E+00	8.021
C9JSB2	0.348	25.827	5.532	0.00E+00	0.00E+00	7.5
M0QZC9	-0.386	25.117	-5.392	0.00E+00	0.00E+00	6.859
MARF1	0.326	25.32	5.121	0.00E+00	0.00E+00	5.653
G3XAK4	-0.308	28.575	-5.019	0.00E+00	0.00E+00	5.209
A0A087WX08	-0.357	26.053	-4.964	0.00E+00	0.00E+00	4.973
H7C1N8	0.335	29.118	4.983	0.00E+00	0.00E+00	5.054
PVRL1	0.318	29.093	4.94	0.00E+00	0.00E+00	4.871

Table 4.3: Significant DEPs in the comparison of proteins in AD vs PSP samples

PID	logFC_sig_3	AveExpr_sig_3	t_sig_3	P.Value_sig_3	adj.P.Val_sig_3	B_sig_3
H7C545	1.422	29.007	14.777	0.00E+00	0.00E+00	66.177
PLEC	0.842	25.819	14.077	0.00E+00	0.00E+00	61.305
ARAID	0.944	37.127	13.075	0.00E+00	0.00E+00	54.322
K7EKD1	1.381	28.46	12.204	0.00E+00	0.00E+00	48.285
H0YD17	2.125	27.558	11.41	0.00E+00	0.00E+00	42.826
ADDG	0.343	31.651	11.41	0.00E+00	0.00E+00	42.829
PADI2	0.696	31.338	11.374	0.00E+00	0.00E+00	42.581
H0Y933	0.479	24.95	11.312	0.00E+00	0.00E+00	42.16
PLEC.1	0.475	35.299	10.999	0.00E+00	0.00E+00	40.029
F2Z2Z8	0.545	29.647	10.911	0.00E+00	0.00E+00	39.433
H0YG30	0.672	27.101	10.882	0.00E+00	0.00E+00	39.241
A0A087X134	0.488	32.848	10.838	0.00E+00	0.00E+00	38.942
AHNK	0.67	31.669	10.823	0.00E+00	0.00E+00	38.841
PLCD1	0.531	29.489	10.629	0.00E+00	0.00E+00	37.532
J3KRM4	0.577	33.036	10.254	0.00E+00	0.00E+00	35.024

2.2.6.2: Inferential statistics and visualization:

Taking the differentially expressed proteins, the following analysis and visualization was done for each pairwise comparisons (AD vs C, PSP vs C, AD vs PSP).

2.2.6.3: Hierarchical clustering and heatmaps:

Hierarchical clustering was performed to check which set of proteins show similar patterns and cluster together among the experimental conditions, using hclust R package followed by the generation of heatmaps. The input was normalized data with proteins in the rows and experimental conditions in the columns. Pheatmap package in R was used to generate the heatmaps. Firstly, Euclidean distance between experiments for AD-Control samples was calculated and then for proteins corresponding to those samples (**Figure 5.1, Figure 5.2**). The data was filtered for AD vs Control samples according to a relaxed threshold of significance ($p < 0.05$). Here, log fold change was set to 0.585 ($\log_2(1.5)$) which represents a 1.5x change. The colour and intensity of the boxes in the heatmap (**Figure 5.3**) are based on the expression changes of protein expressions. Red color represents up-regulated genes and blue represents down-regulated genes. White represents unchanged expression. The same was done for PSP vs Control (**Figure 6**) and AD vs PSP (**Figure 7**)

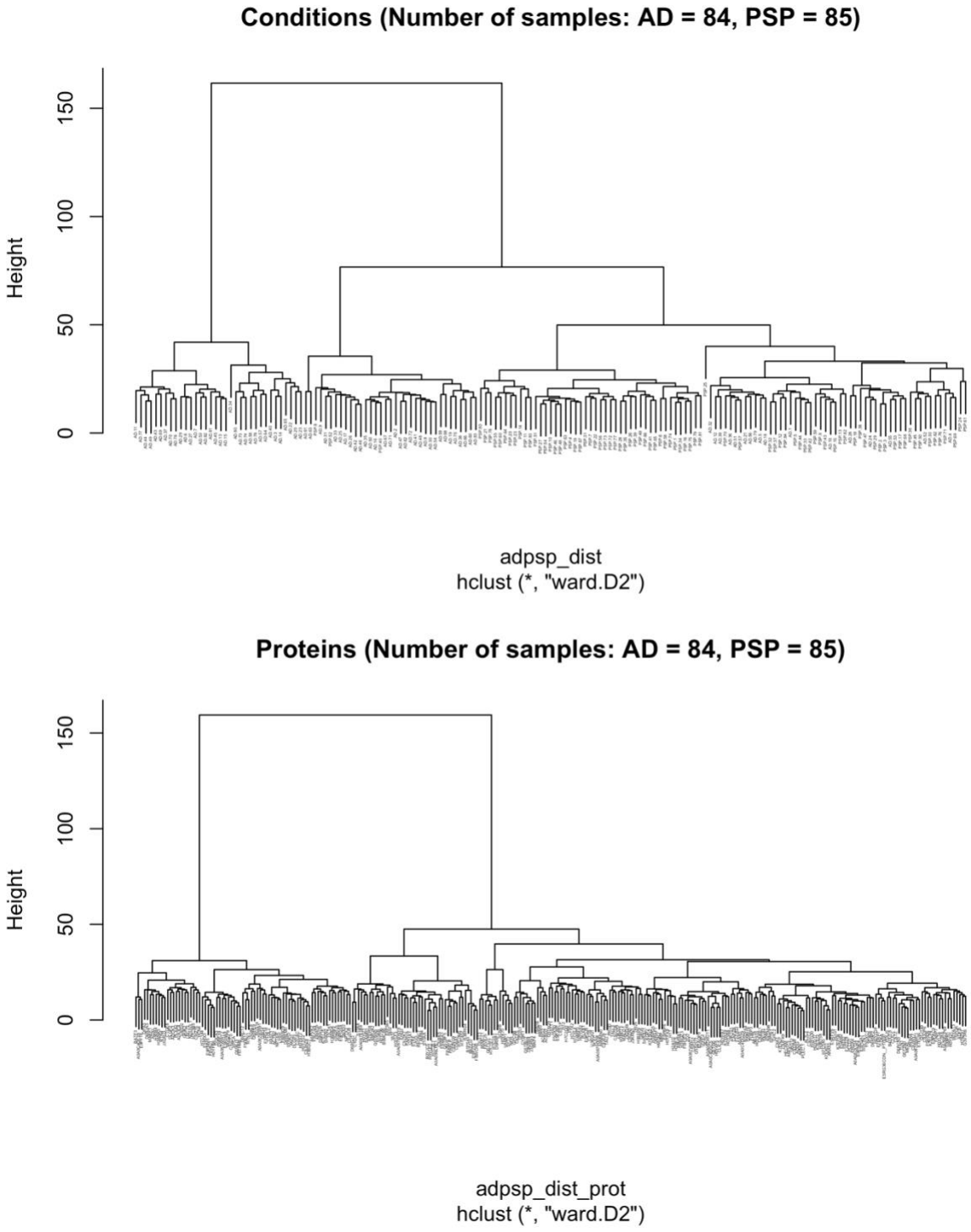
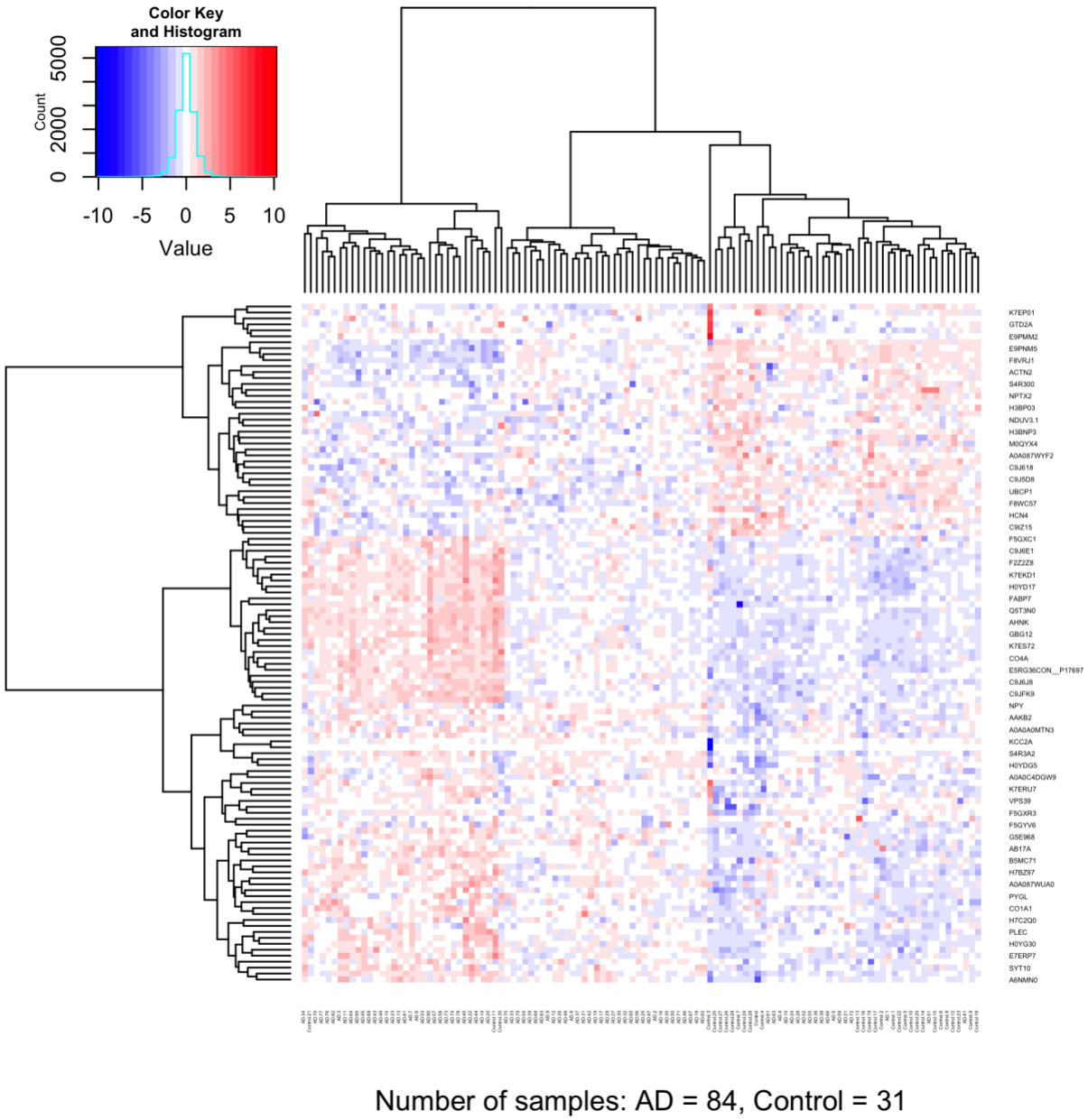


Figure 5.1: Clustering of samples in the experimental conditions for the pairwise comparison in AD vs C and Clustering of proteins in all the samples for the pairwise comparison in AD vs C



Number of samples: AD = 84, Control = 31

Figure 5.2: Hierarchical Clustering and heatmap (AD vs Control): Left: Clustering with respect to samples (above), clustering with respect to proteins (below). Right: The rows represent proteins, and the columns represent AD and Control samples. All upregulated proteins are red in color and downregulated proteins are blue in color. The ones in white are those that show no expression changes.

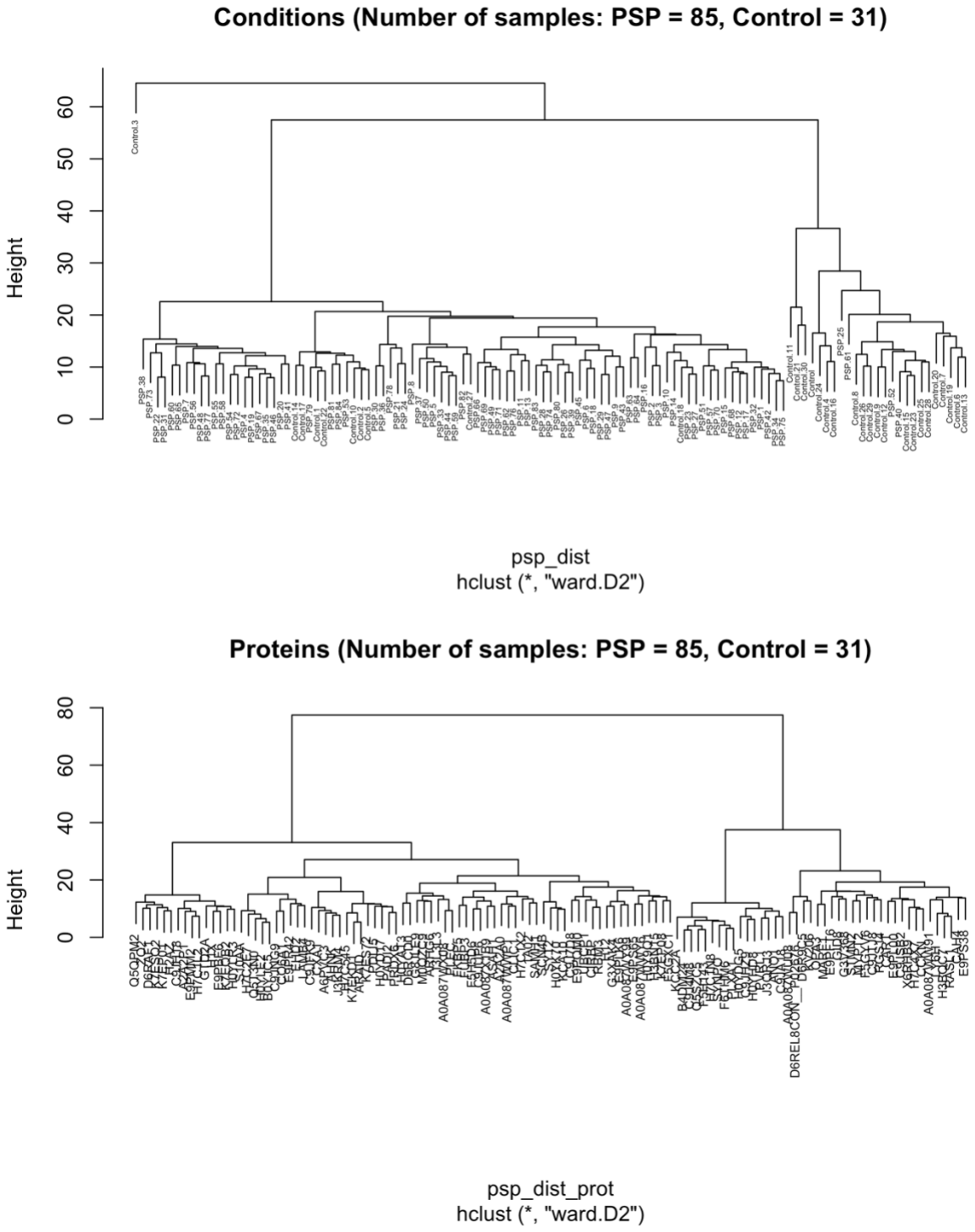


Figure 6.1: Clustering of samples in the experimental conditions for the pairwise comparison in PSP vs C and Clustering of proteins in all the samples for the pairwise comparison in PSP vs C

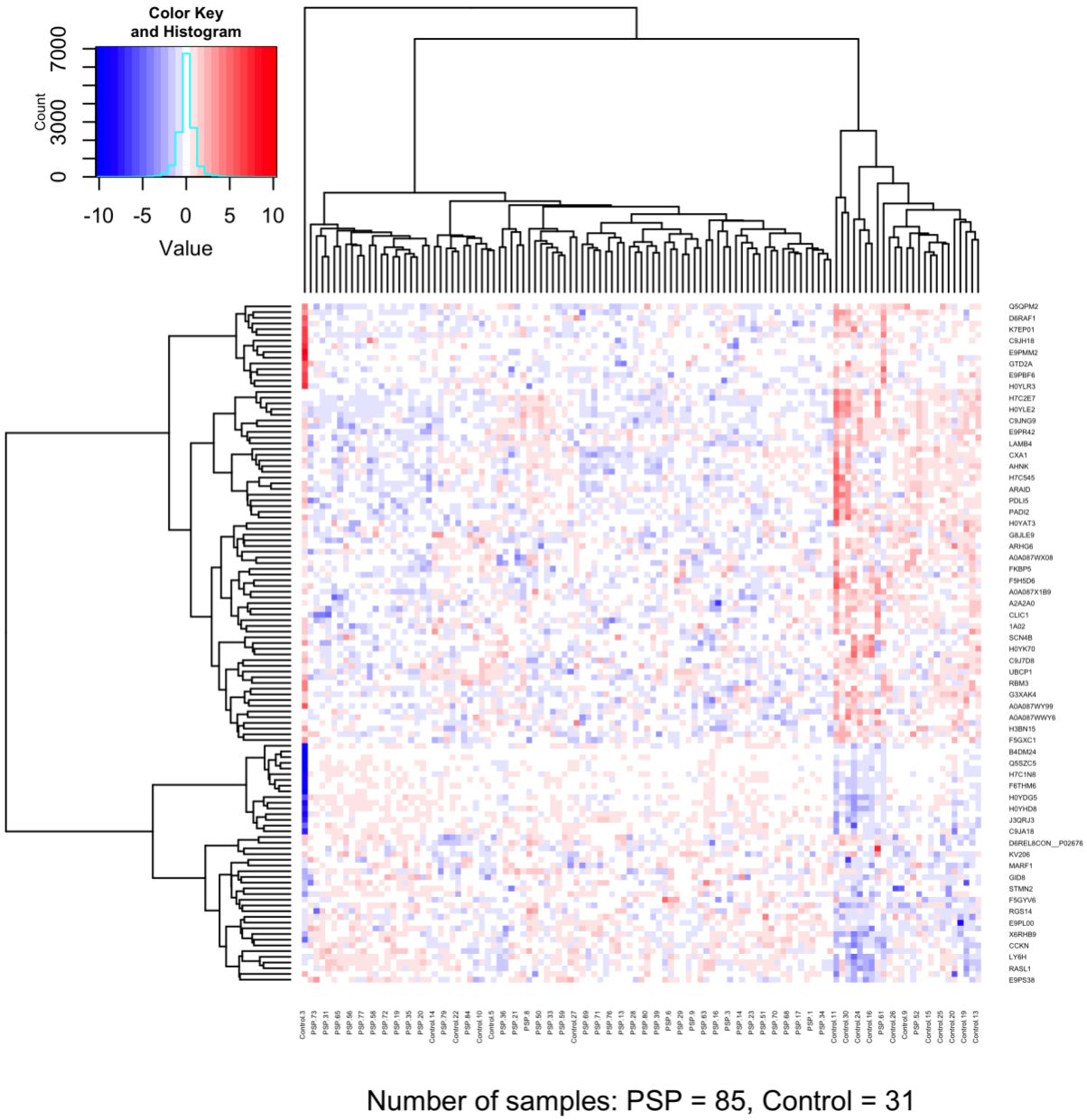


Figure 6.2: Hierarchical Clustering and heatmap (PSP vs C): Left: Clustering with respect to samples (above), clustering with respect to proteins (below). Right: The rows represent proteins, and the columns represent PSP and Control samples. All upregulated proteins are red in color and downregulated proteins are blue in color. The ones in white are those that show no expression changes.

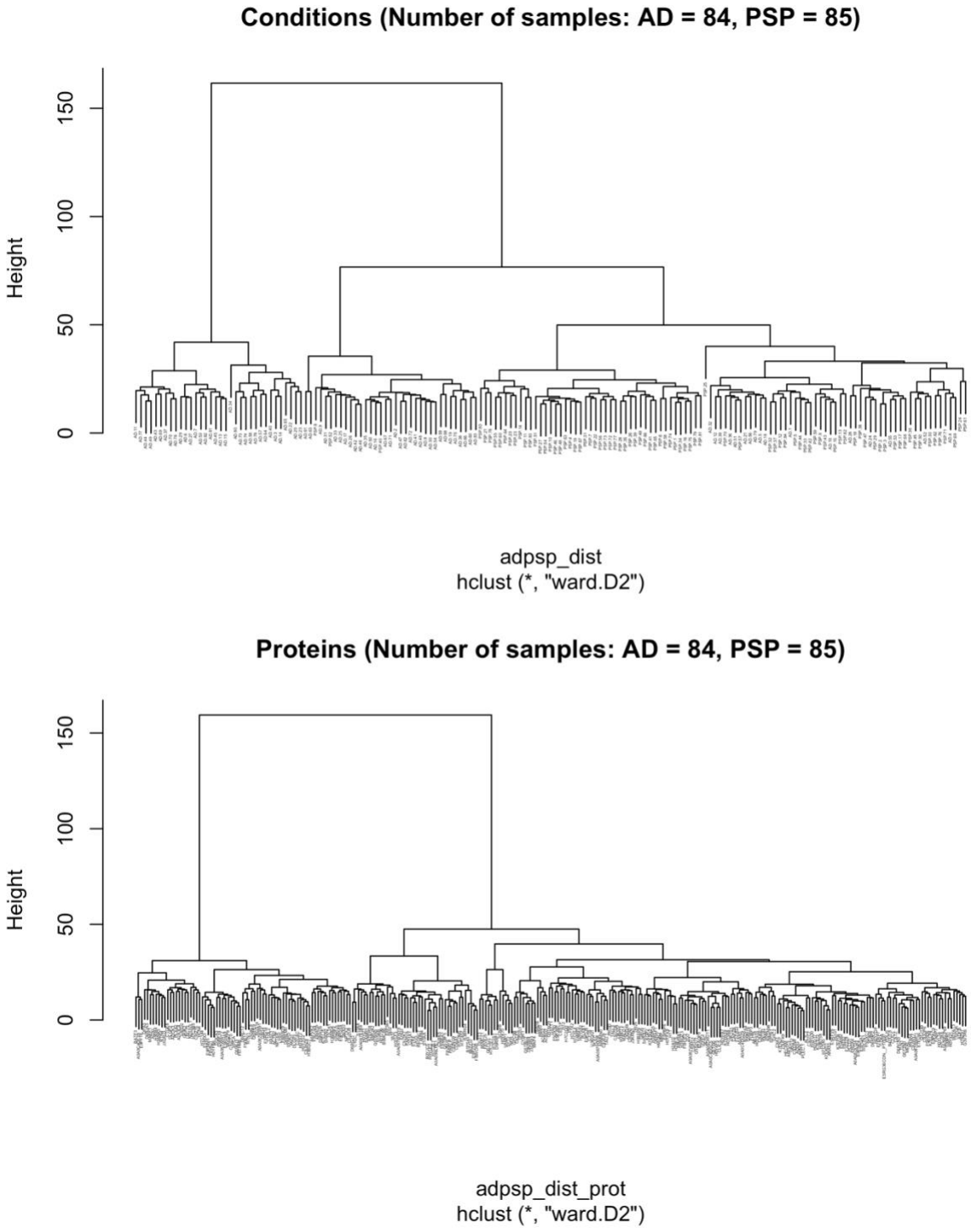


Figure 7.1: Clustering of samples in the experimental conditions for the pairwise comparison in AD vs PSP and Clustering of proteins in all the samples for the pairwise comparison in AD vs PSP

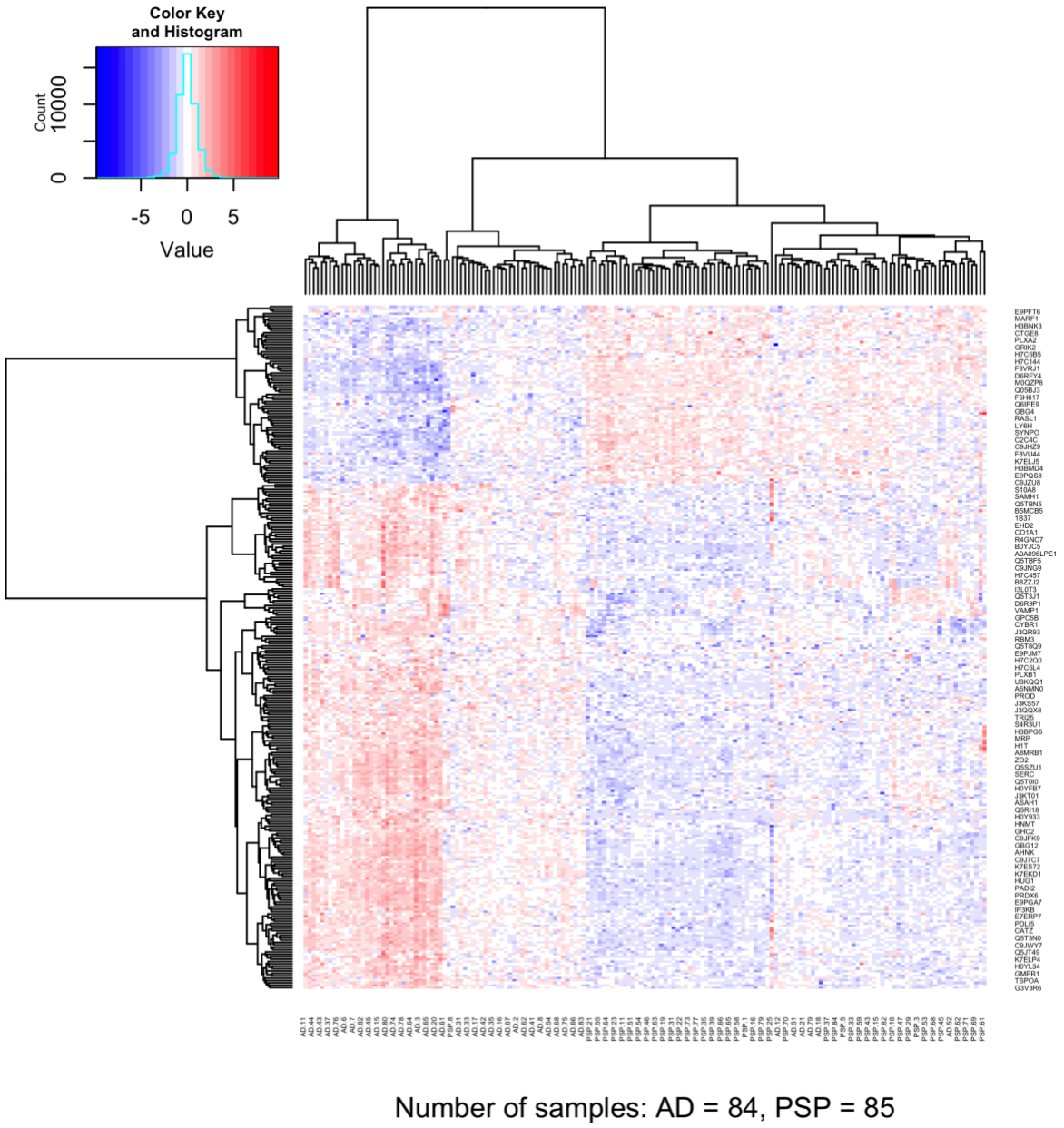


Figure 7.2: Hierarchical Clustering and heatmap (AD vs PSP): left: Clustering with respect to samples (above), clustering with respect to proteins (below). Right: The rows represent proteins, and the columns represent AD and PSP samples. All upregulated proteins are red in color and downregulated proteins are blue in color. The ones in white are those that show no expression changes.

2.2.6.4: Venn diagrams:

To see the extent of overlapping proteins and unique proteins across the three pairwise comparisons, namely AD vs. C, PSP vs. C and PSP vs. AD, a Venn diagram (**Figure 8.1**) was plotted using the Venn diagram package in R with adjusted p-value cut-off adj-p < 0.05. Two other Venn diagrams were plotted to visualize just the upregulated proteins (**Figure 8.2**) and (**Figure 8.3**) in all three pairwise comparisons.

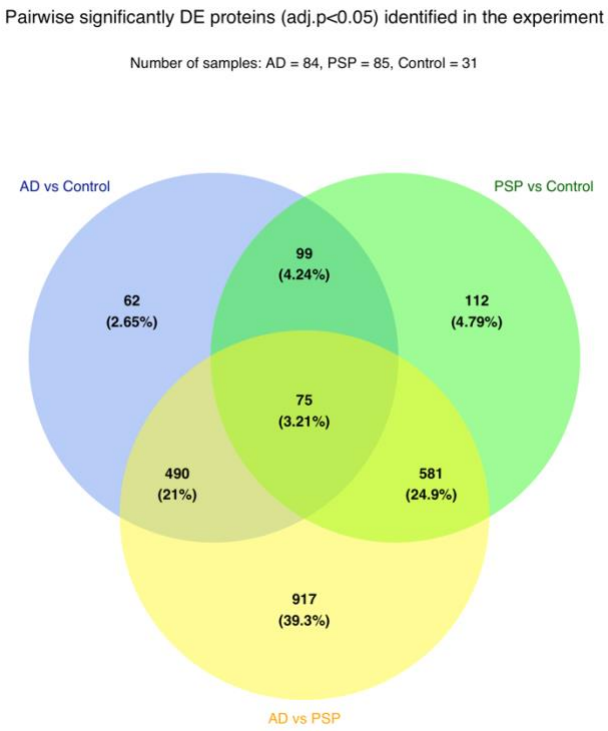


Figure 8.1: Venn diagram showing the overlap of all proteins in the three comparisons. (AD vs C, PSP vs C, AD vs PSP). With respect to AD: The blue represents 43 proteins unique to AD. There are 37 proteins overlapping between AD vs Control and PSP vs Control. 14 proteins are common between the three pairwise comparisons (AD vs C, PSP vs C, AD vs PSP). With respect to PSP: The green represents 27 proteins unique to PSP. With respect to AD and PSP: The yellow represents 1595 proteins unique to AD + PSP.

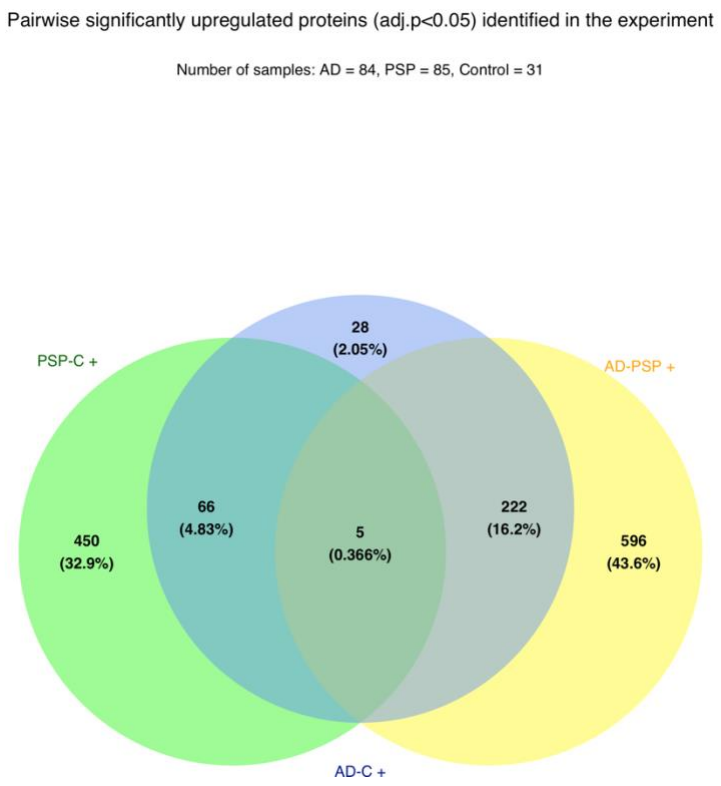


Figure 8.2: Venn diagram showing the overlap of all up regulated proteins in the three comparisons. (AD vs C, PSP vs C, AD vs PSP). With respect to AD: The blue represents 21 unique proteins upregulated in AD. 3 proteins are common between the three pairwise comparisons (AD vs C, PSP vs C, AD vs PSP). With respect to PSP: The green represents 158 proteins unique to PSP. With respect to AD and PSP: The yellow represents 720 proteins unique to AD + PSP.

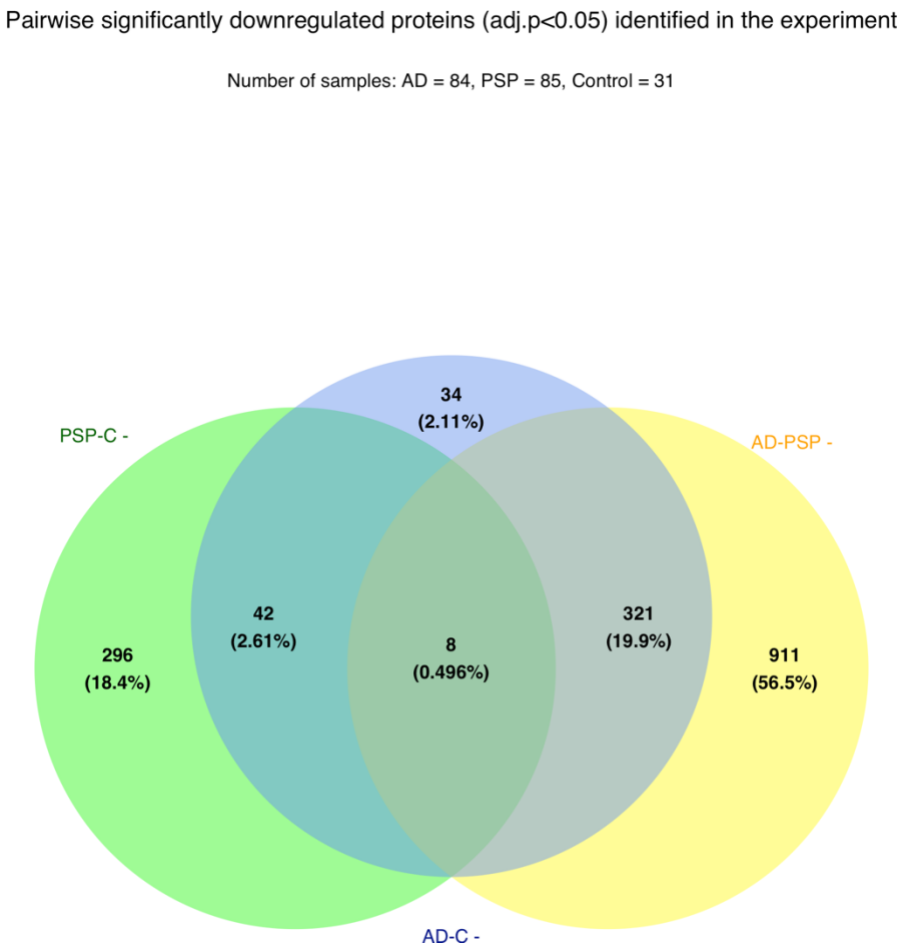


Figure 8.3: Venn diagram showing the overlap of all down regulated proteins in the three comparisons. (AD vs C, PSP vs C, AD vs PSP). With respect to AD: The blue represents 22 unique proteins that are down regulated in AD. 2 proteins are common between the three pairwise comparisons (AD vs C, PSP vs C, AD vs PSP). With respect to PSP: The green represents 71 proteins unique to PSP. With respect to AD and PSP: The yellow represents 1075 proteins unique to AD + PSP.

2.2.6.5: Volcano plots:

Volcano plots were also plotted to visualize up and down regulated proteins across pairwise comparison of the three conditions AD, Control and PSP using EnhanceVolcanoPlot, an R package. The significant proteins (with $P < 0.05$ and log2fc >=0.3 and >=-0.3) are highlighted in red. Those highlighted in black are the proteins with non-significant differences in expression levels, and those highlighted in green are proteins with log2fc >= 0.3 and >= -0.3 in all the pairwise comparisons AD vs. C, PSP vs. C and AD vs. PSP.

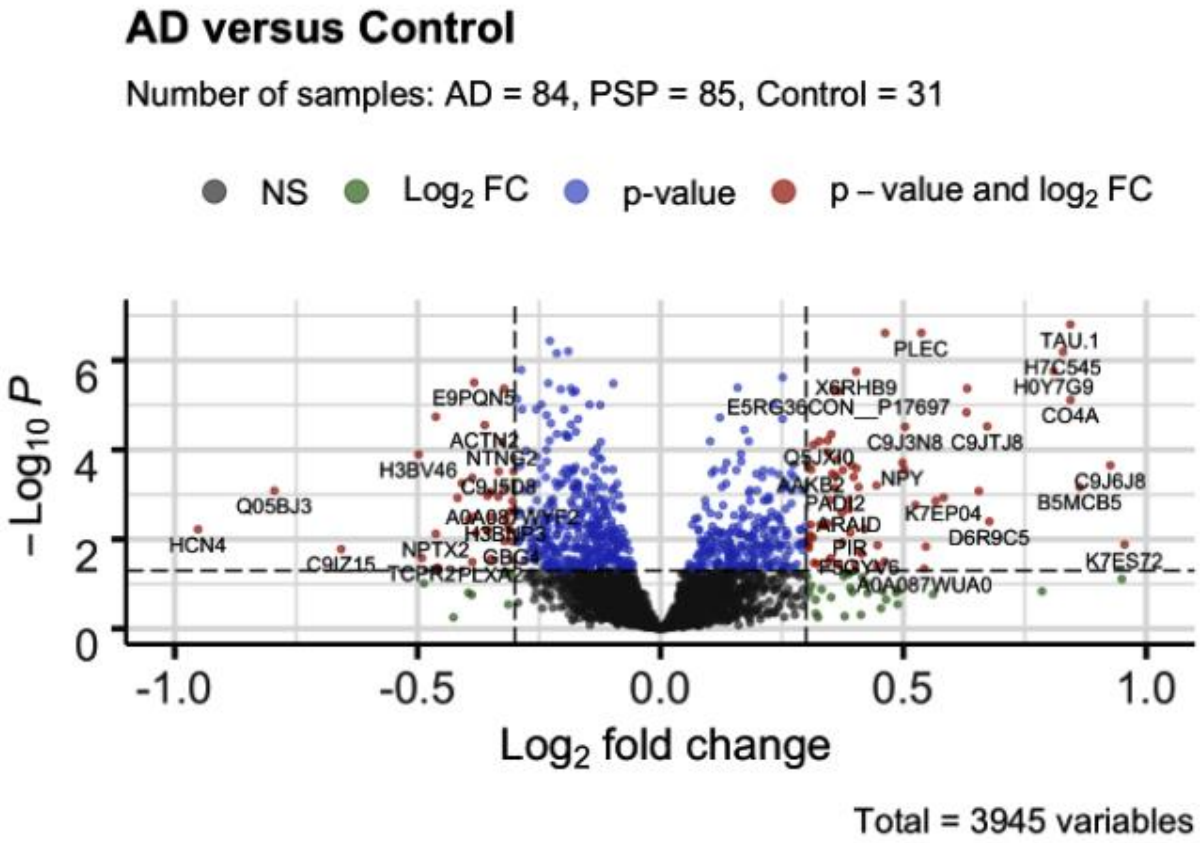


Figure 9.1: Volcano plot showing all dysregulated proteins in AD vs C. The significant proteins (with $P < 0.05$ and log2fc >=0.3 and >=-0.3) are highlighted in red. Black = proteins with non-significant differences in expression levels, green = proteins with log2fc >=0.3 and >=-0.3.

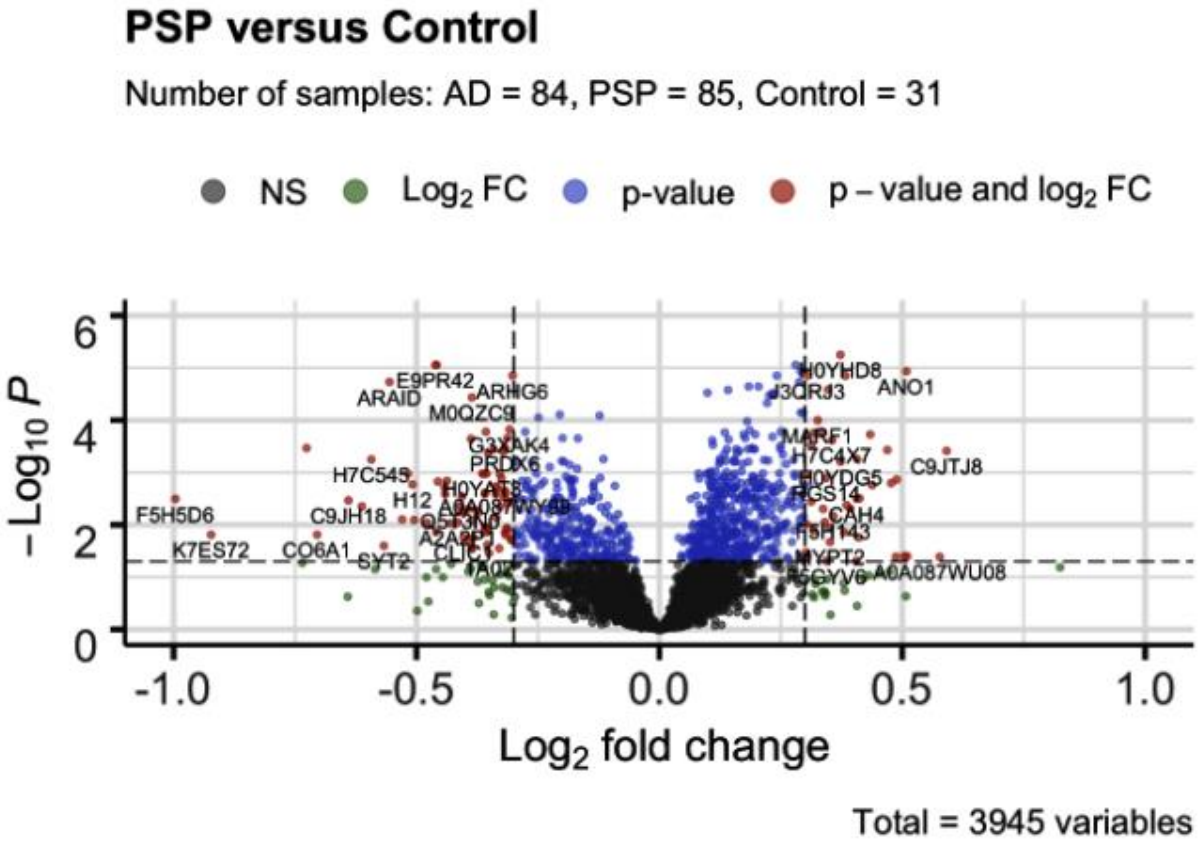


Figure 9.2: Volcano plot showing all dysregulated proteins in PSP vs C. The significant proteins (with $P < 0.05$ and log2fc >=0.3 and >=-0.3) are highlighted in red. Black = proteins with non-significant differences in expression levels, green = proteins with log2fc >=0.3 and >=-0.3.

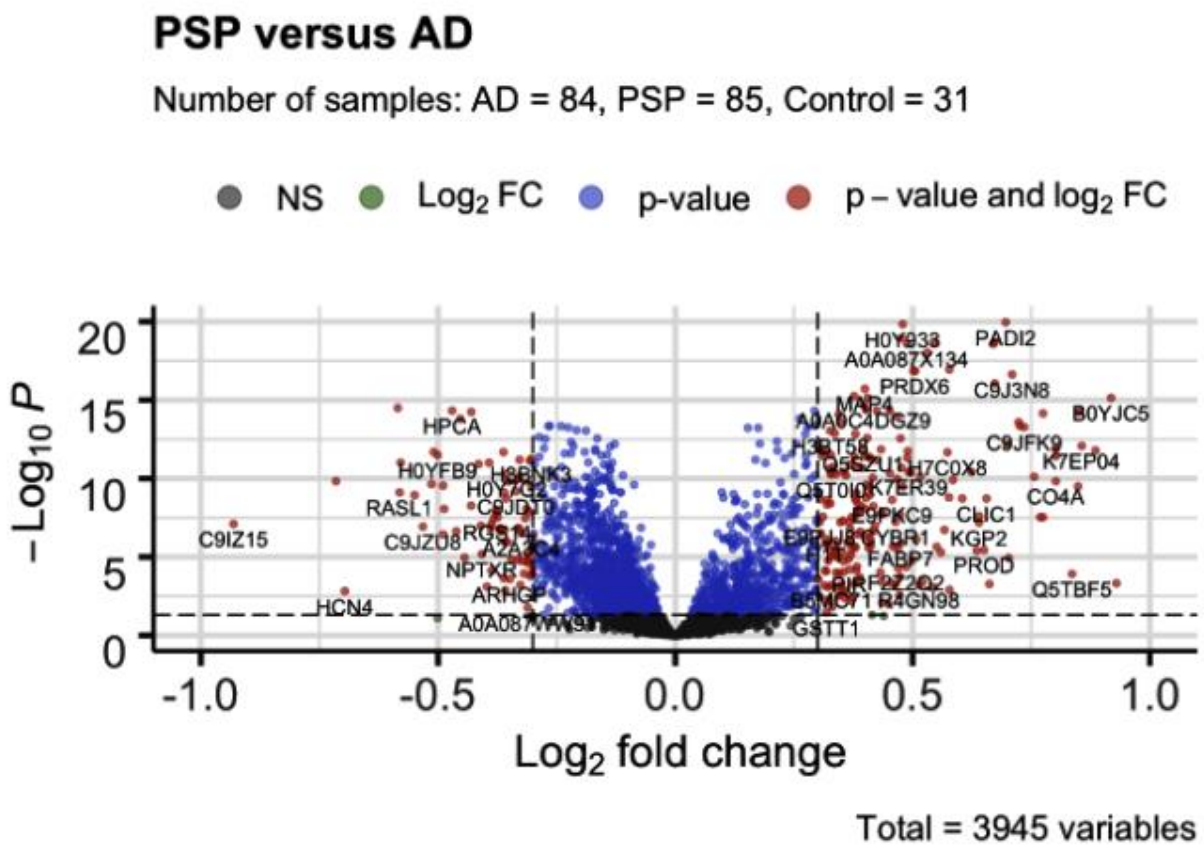


Figure 9.3: Volcano plot showing all dysregulated proteins in PSP and AD. The significant proteins (with $P < 0.05$ and log2fc >=0.3 and >=-0.3) are highlighted in red. Black = proteins with non-significant differences in expression levels, green = proteins with log2fc >=0.3 and >=-0.3.

2.2.6.6: Box Plots:

We wanted to check the distribution of protein expression for each condition. So, we plotted boxplots (AD vs C, PSP vs C, AD vs PSP). The x-axis represents samples for all pairwise comparisons (AD and Control, PSP and Control, AD and PSP), and y-axis represents their protein expression. We observed that for proteins C9JTJ8, HCN4 and CO1A1, the protein expression levels are high compared to other proteins in AD as can be seen in **Figure 10.1**.

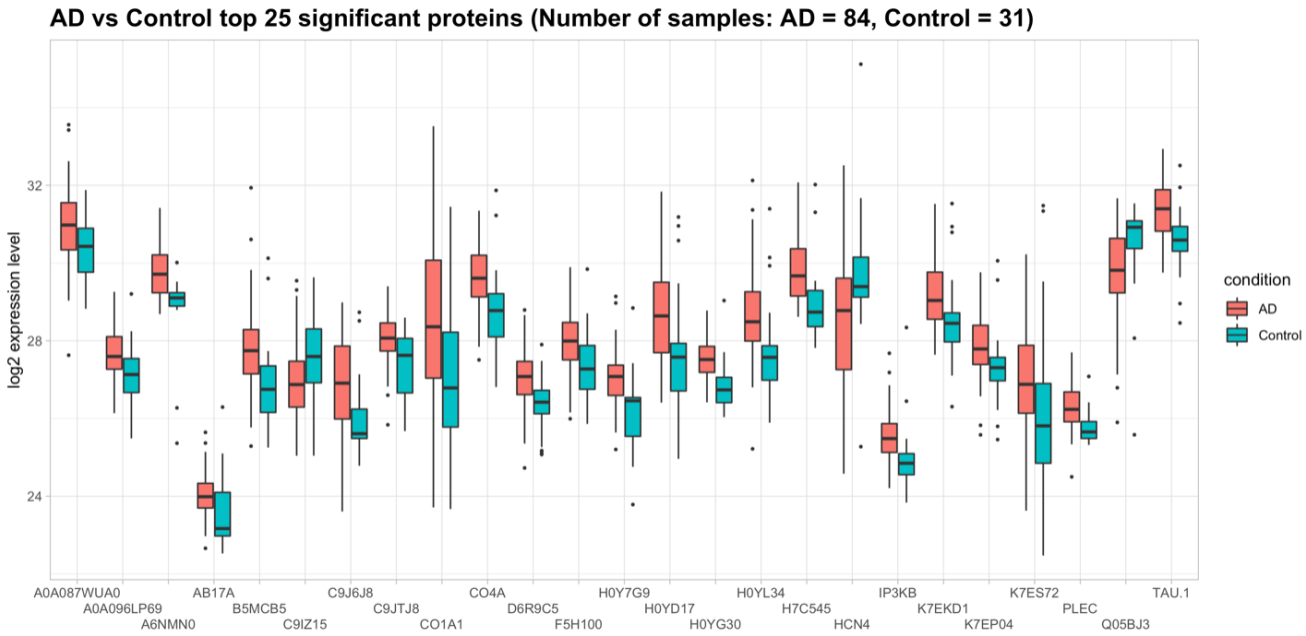


Figure 10.1: Boxplots for AD vs C. The protein expression levels are high for proteins C9JTJ8, HCN4 and CO1A1 in AD.

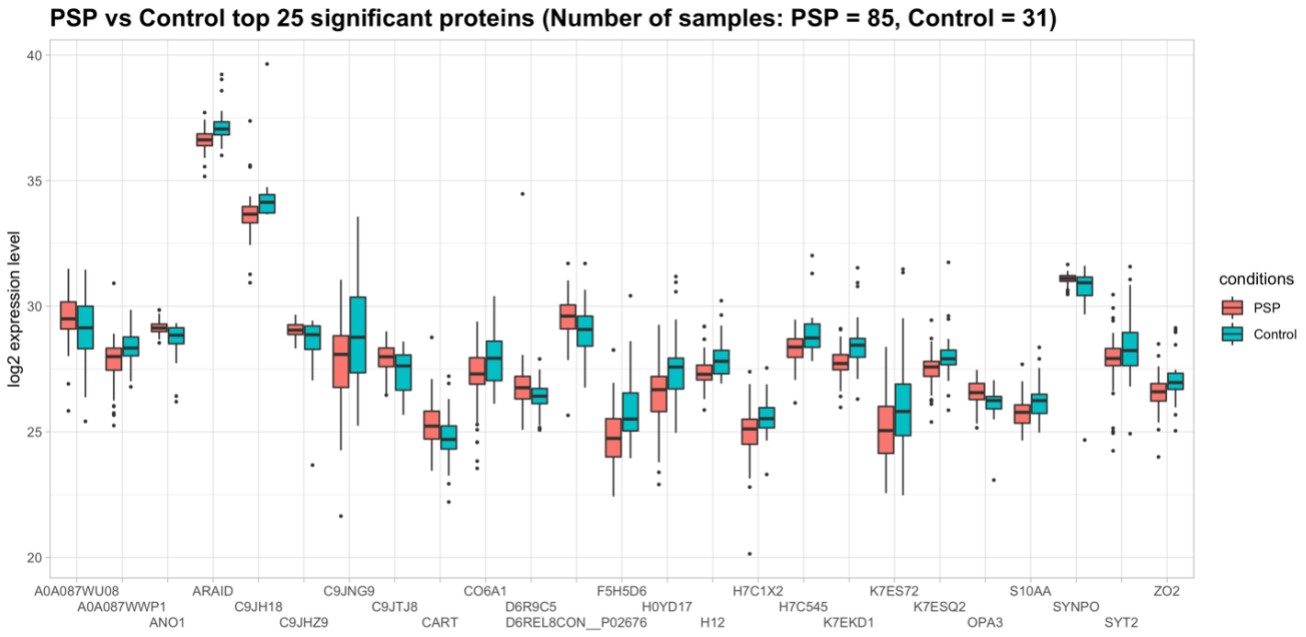


Figure 10.2: Boxplots for PSP vs C The protein expression levels are high for C9JNG9, F5HD6, H0YD17, K7ES72 in PSP.

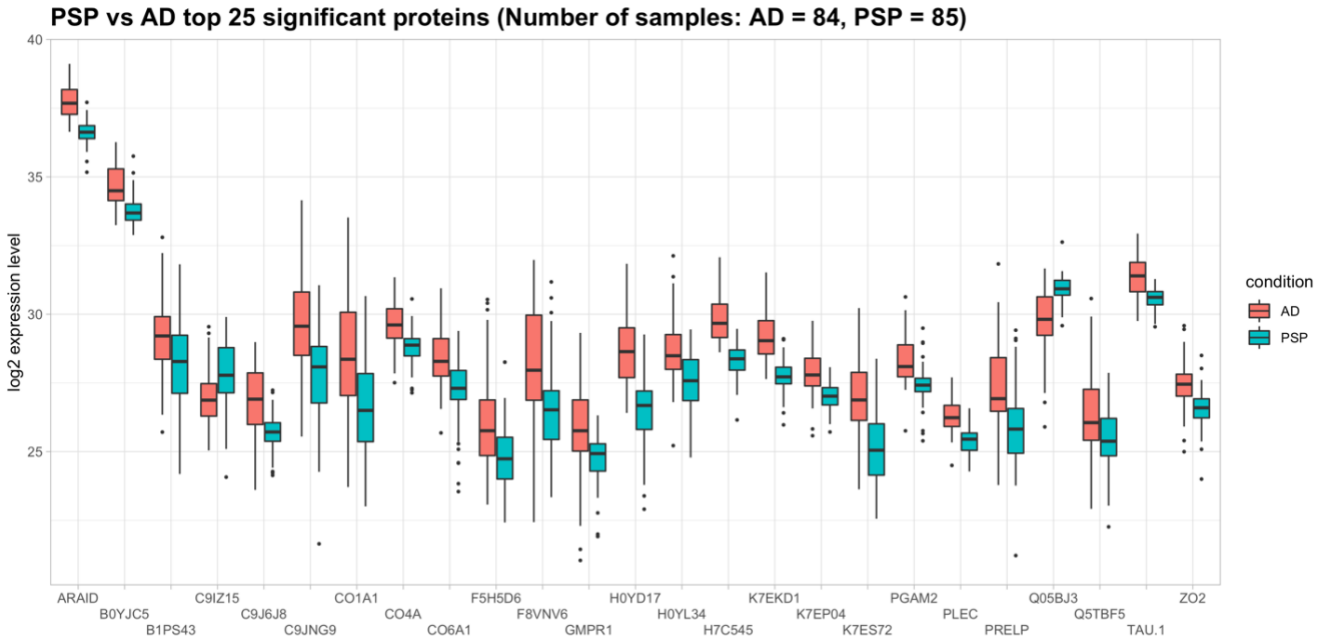


Figure 10.3: Boxplots for AD vs PSP. The protein expression levels are high for C9JNG9, F5HD6, H0YD17, PRELP.

2.2.6.7: Enrichment analysis of pathways and process:

After performing a DEP analysis and obtaining significant DEPs, they were used for functional enrichment analysis of the pairwise comparison AD vs. C. The functional enrichment analysis consisted of KEGG pathways, Gene Ontology process. Protein-protein interaction networks were constructed.

2.2.6.7.2: Gene Ontology and KEGG pathways:

To look at the common pathways and cellular processes shared by the proteins of interest, KEGG pathway analysis and Gene Ontology were performed.

KEGG pathways: KEGG pathways tell us about various biochemical processes like metabolism, genetic and environmental information processing, cellular processes, organismal systems, human diseases.

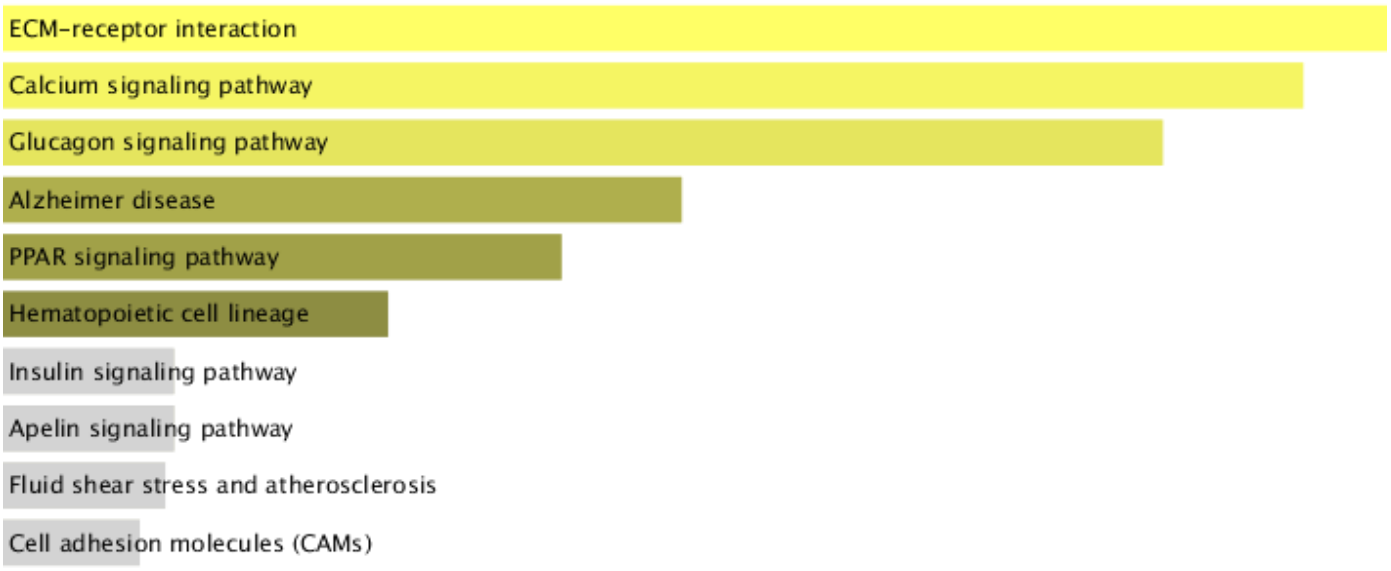


Figure 11.1: KEGG pathways in AD vs. C

Gene Ontology: Gene Ontology provides information about gene functions and relationships between the functions along with three aspects, namely Biological Processes (larger processes of the genes), Molecular Function (molecular activities of the gene), and Cellular Component (which part of the cell, the gene products are active).

a. GO: BP

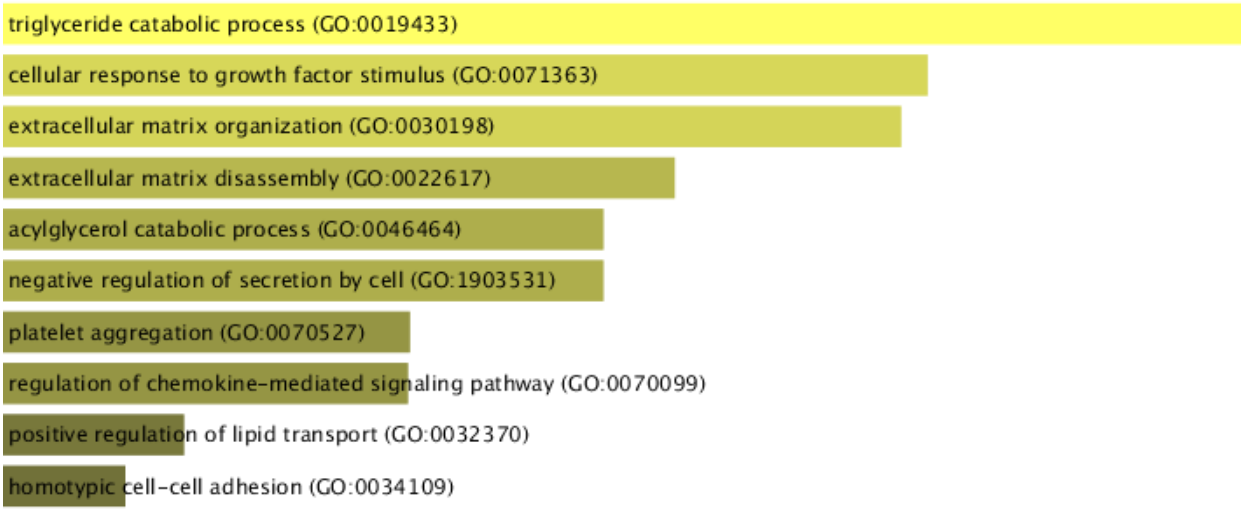


Figure 11.2: Gene Ontology of Molecular Functions in AD vs C

b. GO: MF

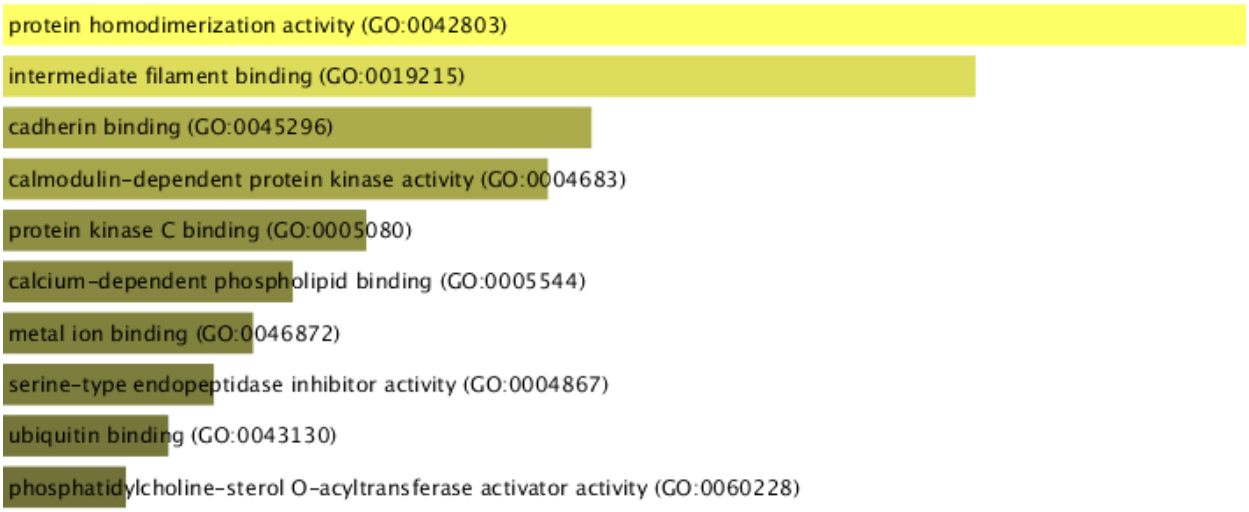


Figure 11.3: Gene Ontology of Molecular Functions in AD vs C

c. GO: CC

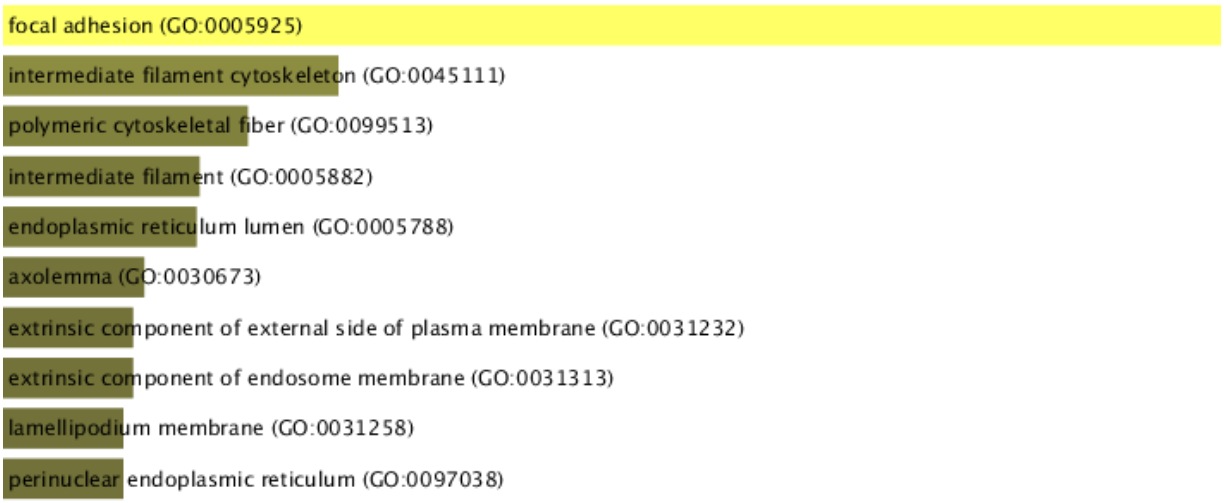


Figure 11.4: Gene Ontology of Cellular Component in AD vs. C

2.2.6.7.3: Protein-Protein Interactions for hub proteins:

It is essential to study proteins and their functional interactions to understand the complete biological phenomenon within a cell. So, we performed a Protein-protein interaction (PPI) enrichment analysis in STRING DB among the list of proteins that were significantly dysregulated to identify the group of proteins that have a similar function (functional association refers to a link between two proteins that both contribute jointly to a specific biological process) and belong to the same pathway rather than looking at the physical interactions. The interaction scores in STRING do not represent the strength or specificity of a given interaction. Instead, they are meant to express approximate confidence on a scale of zero to one. The scores in STRING are benchmarked using the subset of associations for which both protein partners are already functionally annotated (STRING v11- Szklarczyk et al., 2019). So, the hub proteins are those that are functionally interacting with other proteins.

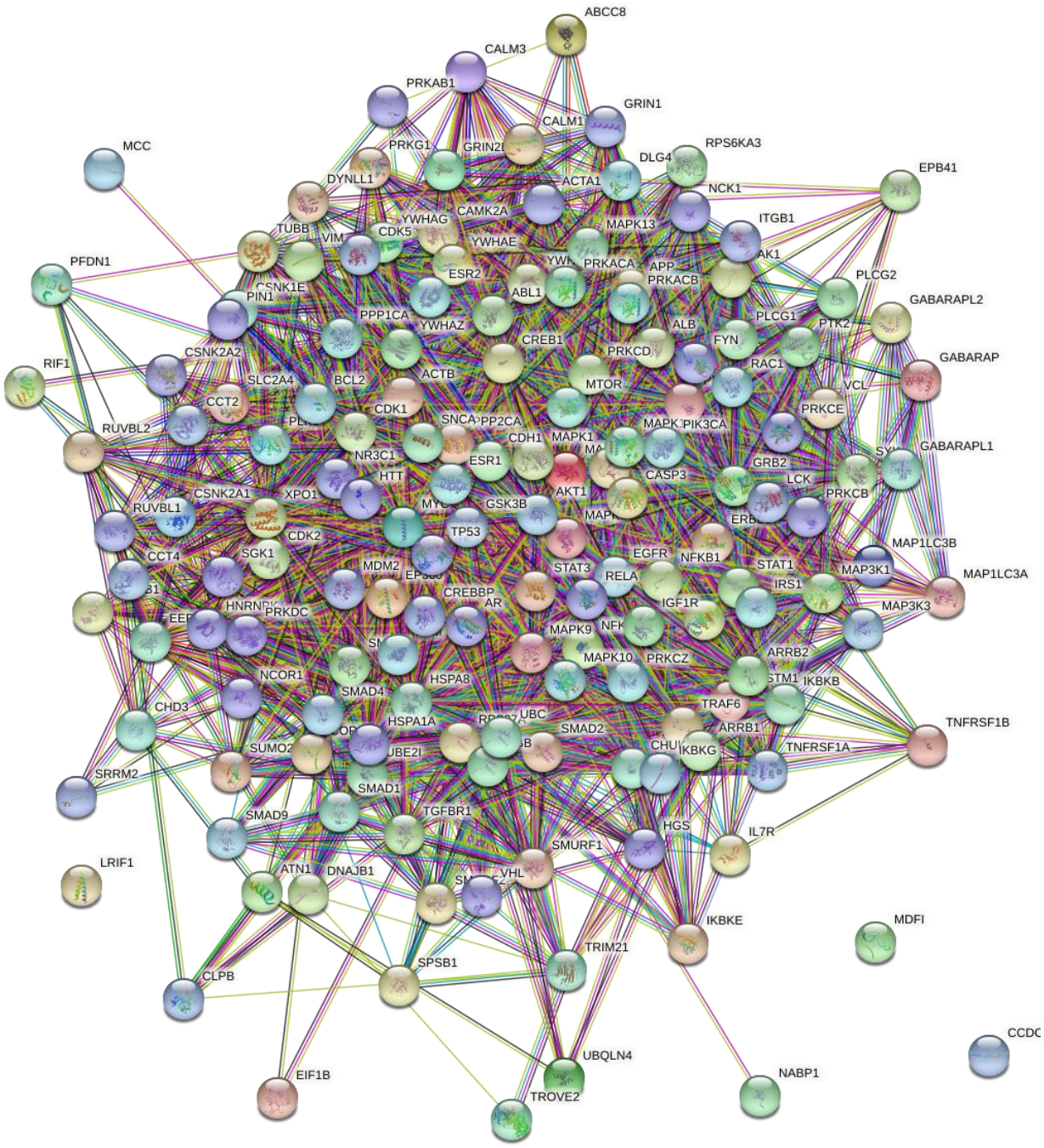


Figure 11.5: Protein-Protein Interactions (PPIs) in AD vs. C

2.3: Differential expression analysis of RNA seq:

We performed the differential expression analysis of RNAseq transcript counts (gene profiling) data following the tutorial https://bioinformatics-core-shared-training.github.io/RNAseq-R/. The data used for the analysis was downloaded from synapse.org (synapseID: syn20818651) named "Mayo_RNAseq_TCX_transcriptCounts.tsv". We chose to process transcript counts for further analysis.

2.3.1: Quality Control (QC):

Transcripts with very low counts across all libraries provide little evidence for differential expression and so have to be filtered out. To filter out lowly expressed genes, we filtered on a minimum counts per million (CPM) threshold present in at least 31 samples. 31 represents the smallest sample size (in the Control group) for each group in our experiment. We choose to retain genes if they are expressed at a CPM above 1 in at least 31 samples. We used the `cpm` function from the *edgeR* library to generate the CPM values. By converting to CPMs we normalized for the different sequencing depths for each sample. Next, we created a `DGEList` object. (This is an object used by *edgeR* to store count data.) To perform quality control, we looked at a few different plots to check that the data quality is good.

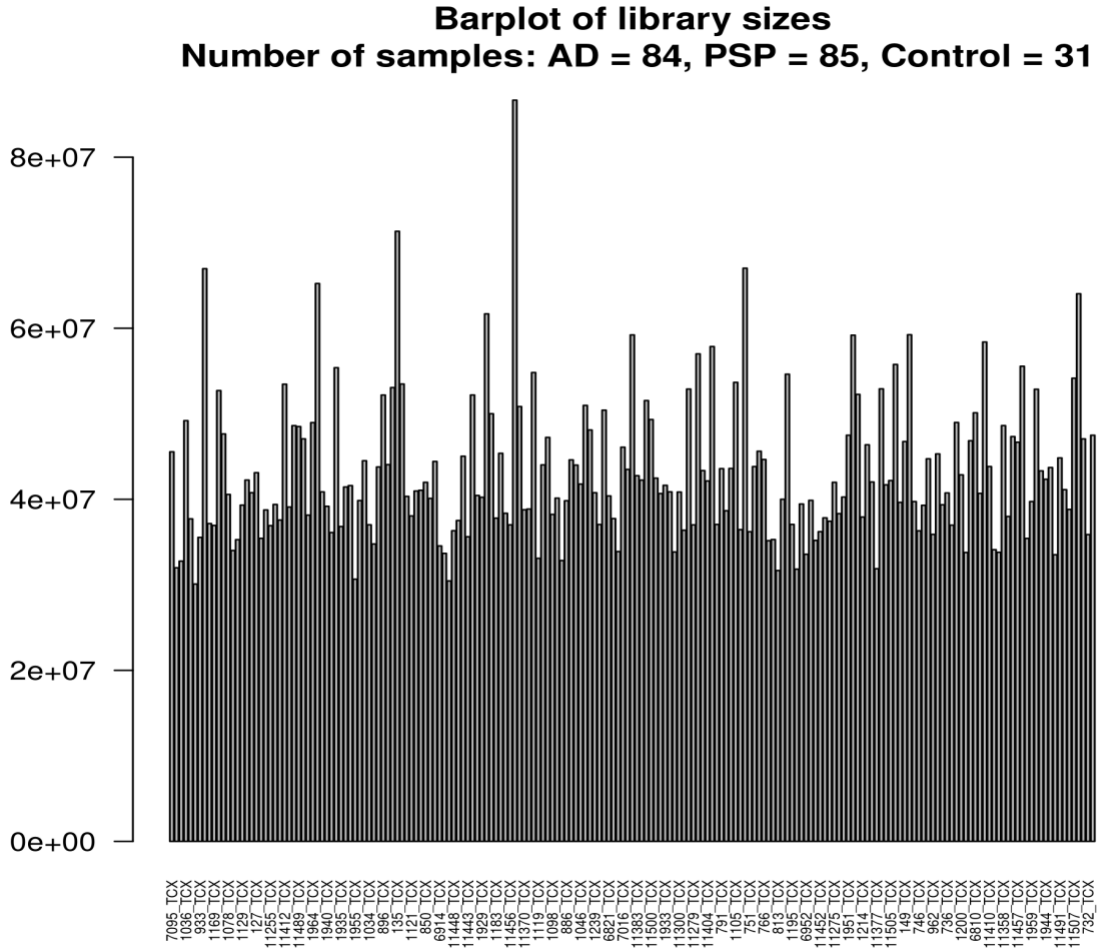


Figure 12: Bar plot of library sizes: Plots of the library sizes as a bar plot to see whether there are any major discrepancies between the samples more easily.

From the plot, we can see that Count data is not normally distributed, so if we want to examine the distributions of the raw counts, we need to log the counts. Next, we used box plots to check the distribution of the read counts on the log2 scale. We can use the `cpm` function to get log2 counts per million, which are corrected for the different library sizes. The `cpm` function also adds a small offset to avoid taking log of zero.

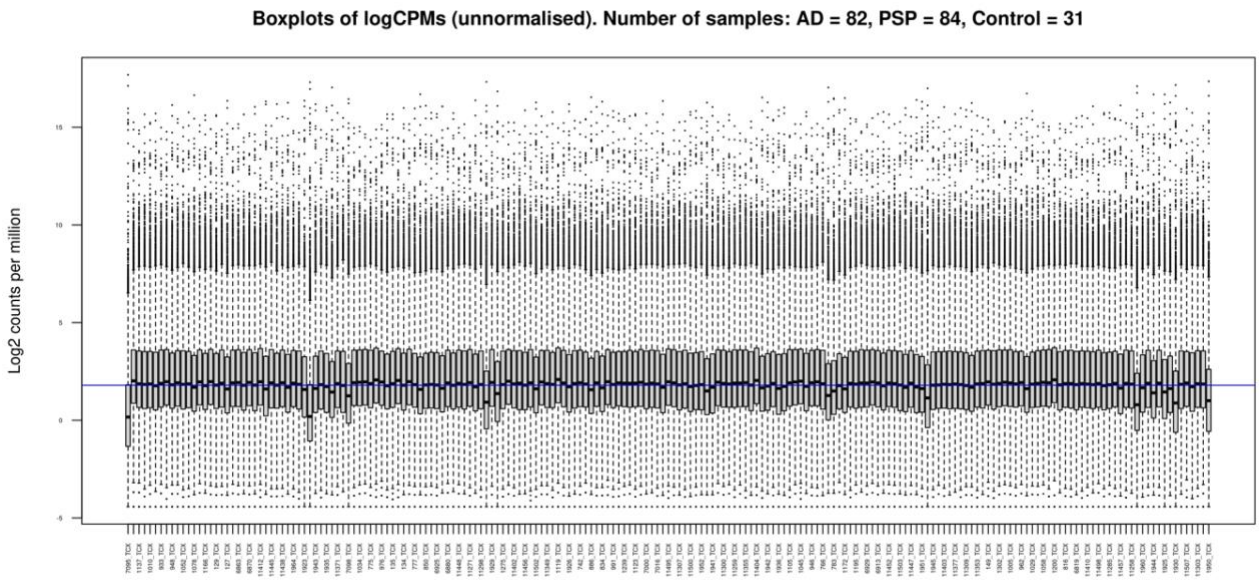


Figure 13: Box plots to check the distribution of the read counts on the log2 scale.

From the boxplots we see that overall, the density distributions of raw log-intensities are not very identical but still not very different. We used the plotMDS function to create the multidimensional (MDS) plot. We color-coded the samples according to the grouping information.

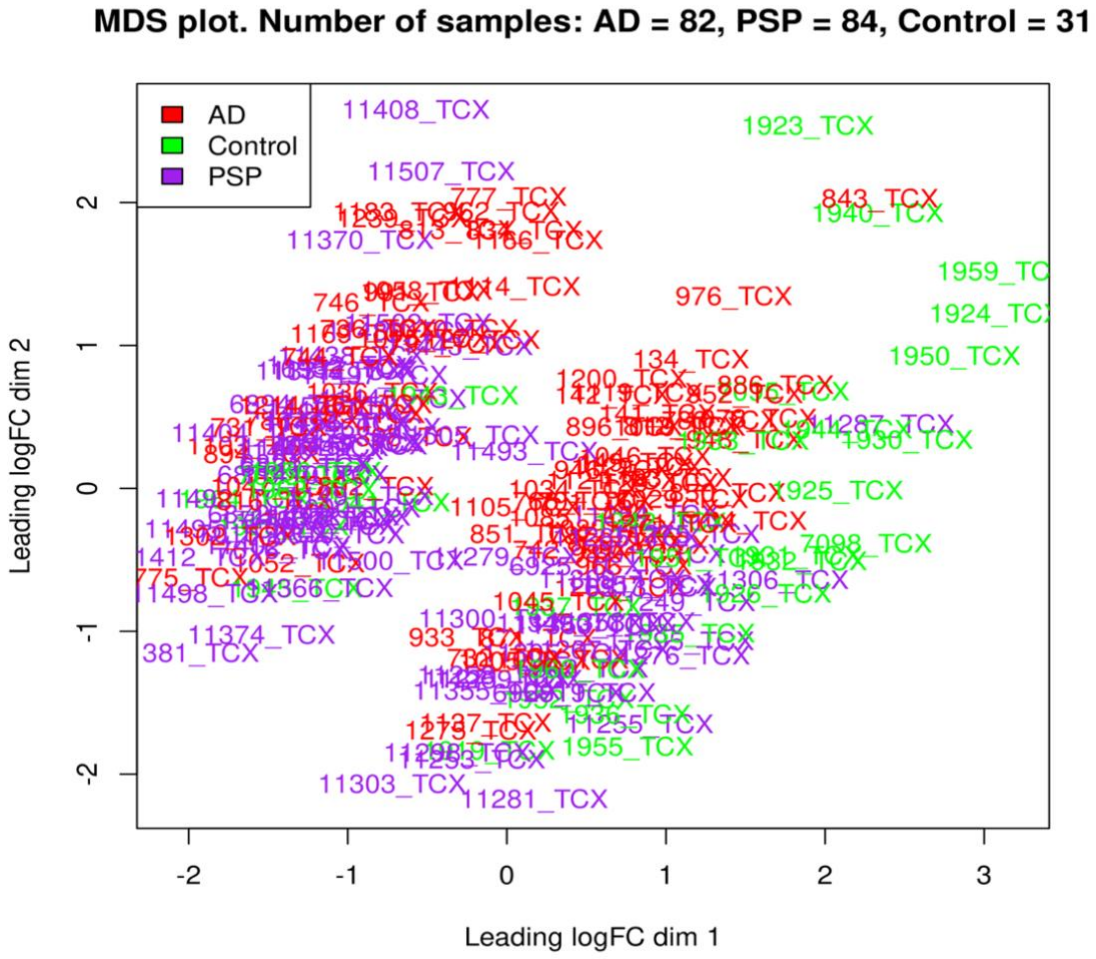


Figure 14: MDS Plot: We used the plotMDS function to create the multidimensional (MDS) plot. We color-coded the samples according to the grouping information.

2.3.2: Normalization:

After checking the quality of our data, we normalized the data using trimmed mean of M-values normalization method (TMM) embedded in edgeR, to eliminate composition biases between libraries. The `calcNormFactors` function in `edgeR` calculates the normalization factors between libraries. The first two samples had very different normalization factors. We plotted mean difference plots using

the `plotMD` function for these samples, we saw the composition bias problem. We used the `logcounts`, which have been normalised for library size, but not for composition bias. So, now we had to work on composition bias.

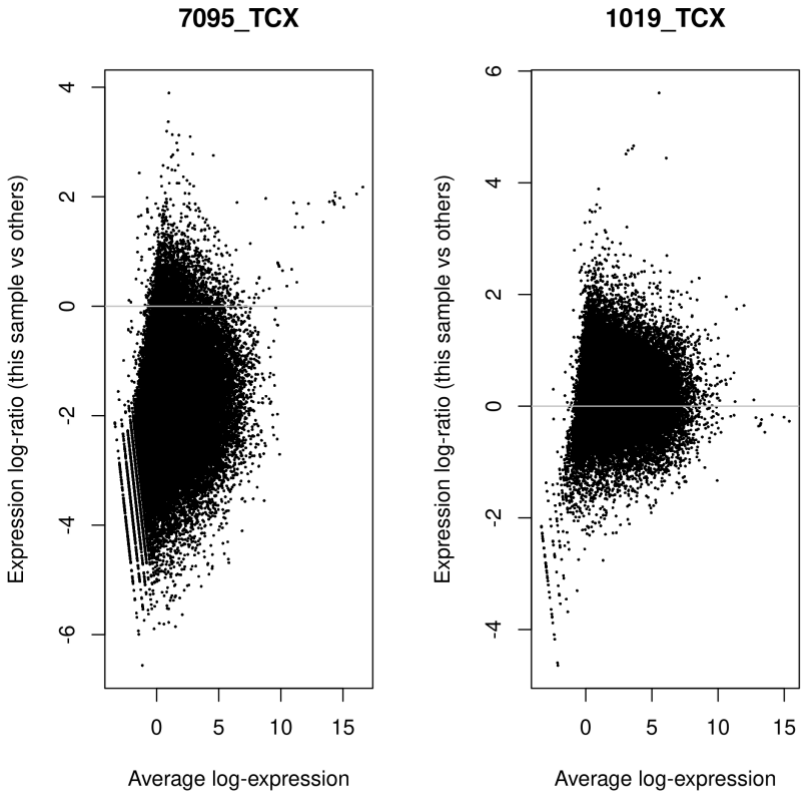


Figure 15a: The mean-difference plots show average expression (mean: x-axis) against log-fold-changes (difference: y-axis).

Since our `DGEList` object contains the normalization factors, plots using `dgeObj (normalized DEGList`, show the composition bias problem has been solved. We can see in the next plot.

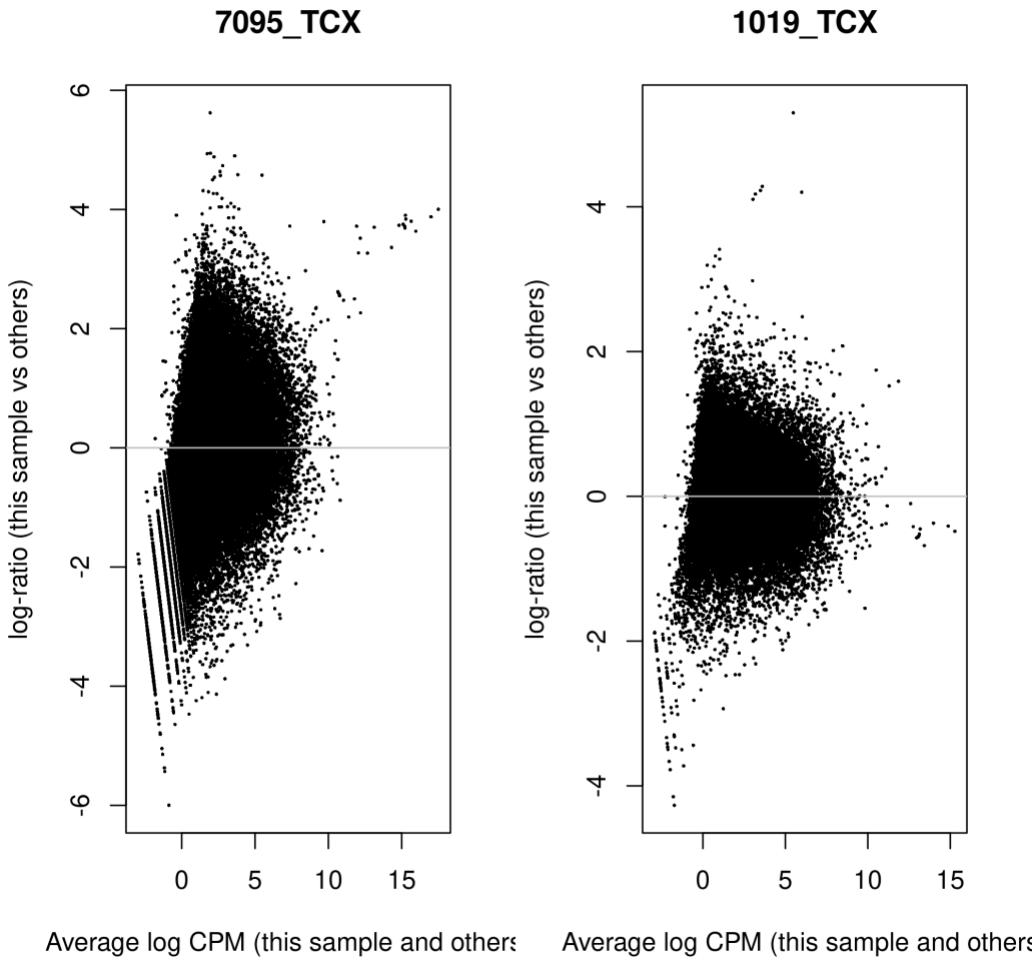


Figure 15b: The normalized mean-difference plots show average expression (mean: x-axis) against log-fold-changes (difference: y-axis).

After normalization, we performed the differential expression analysis between AD and Control, PSP and Control, AD and PSP. We kept the same p-value and fold-change cut offs as in proteomics data analysis. (adj.P-value < 0.05, logFC = 0.03). The list of genes are attached as an appendix. As a part of inferential statistical analysis, to visualize how the sample and gene clusters are grouping themselves, we performed clustering and visualized our samples with heatmaps. We plotted Venn diagrams to check the overlap of common genes and to see how many genes were unique to a particular class of comparisons. We plotted volcano plots to visualize significantly differentially expressed genes. We

also plotted box plots to check the distribution of genes abundances in each condition in comparison with control (Attached in Appendix).

2.4: Correlation analysis:

Correlation analysis is a statistical method of looking at the strength of pairing between two variables. The central dogma of molecular biology has laid in simple and direct terms that DNA makes RNA makes proteins but there is substantial precariousness to it. Correlations are found to exist between the levels of RNA and their corresponding proteins. However, we wanted to see if there existed any correlations between genes from a data set (RNAseq) and proteins from another dataset (proteomics). So, alongside looking at the correlations within each -omics layer, we wanted to check if there existed any correlations between the two omics layers using Spearman's correlation coefficients method. We used Spearman's correlation coefficients because they are rank-based and are robust for logarithmic data in contrast to Pearson's correlation coefficients which could be strongly biased by extreme values [30]. Firstly, we wanted to see if there are any interactions within proteins from the proteomics data set. Secondly, we looked at the interactions within genes in the RNAseq data set. Finally, we looked at correlations between proteins that were obtained from proteomics and genes obtained from RNAseq. We obtained our input data for correlation analysis from the individual differential expression analysis of proteins and genes. For example, to get the set of proteins involved in AD, we selected from the normalized data and from the normalized and log2-transformed protein data only those samples that are also in the RNA data. In the proteomics dataset, 1:31 were Control columns, 32:115 were AD columns and 116:200 were PSP columns (where columns are samples). We removed three samples that were absent in the transcriptomics data so as to match the number of samples from both -omics layers.

Thus, we have equal sample size.

```
Proteins: Number of samples: AD = 82, PSP = 84, Control = 31
Genes: Number of samples: AD = 82, PSP = 84, Control = 31
```

2.4.1: Correlation analysis of proteins involved in AD with proteins of all samples:

We selected AD proteins (N = 114) obtained from AD vs. Control and looked for them in the complete data consisting of AD, Control, and PSP samples.

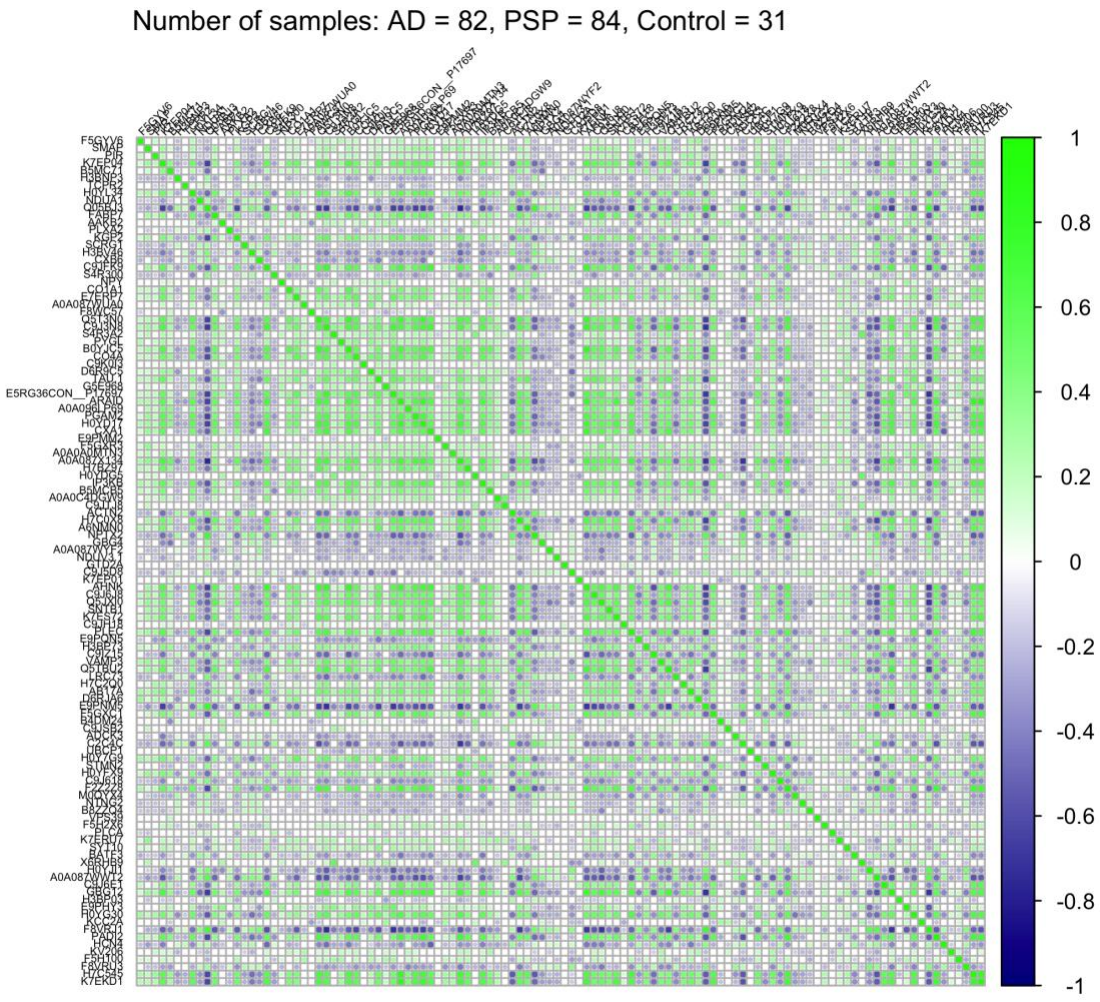


Figure 16.1: Correlation plots of proteins involved in AD with proteins of all samples

2.4.2: Correlation analysis of genes involved in AD with genes in all samples:

We selected AD genes (N = 61) obtained from AD vs. Control and looked for them in the complete data consisting of AD, Control, and PSP samples.

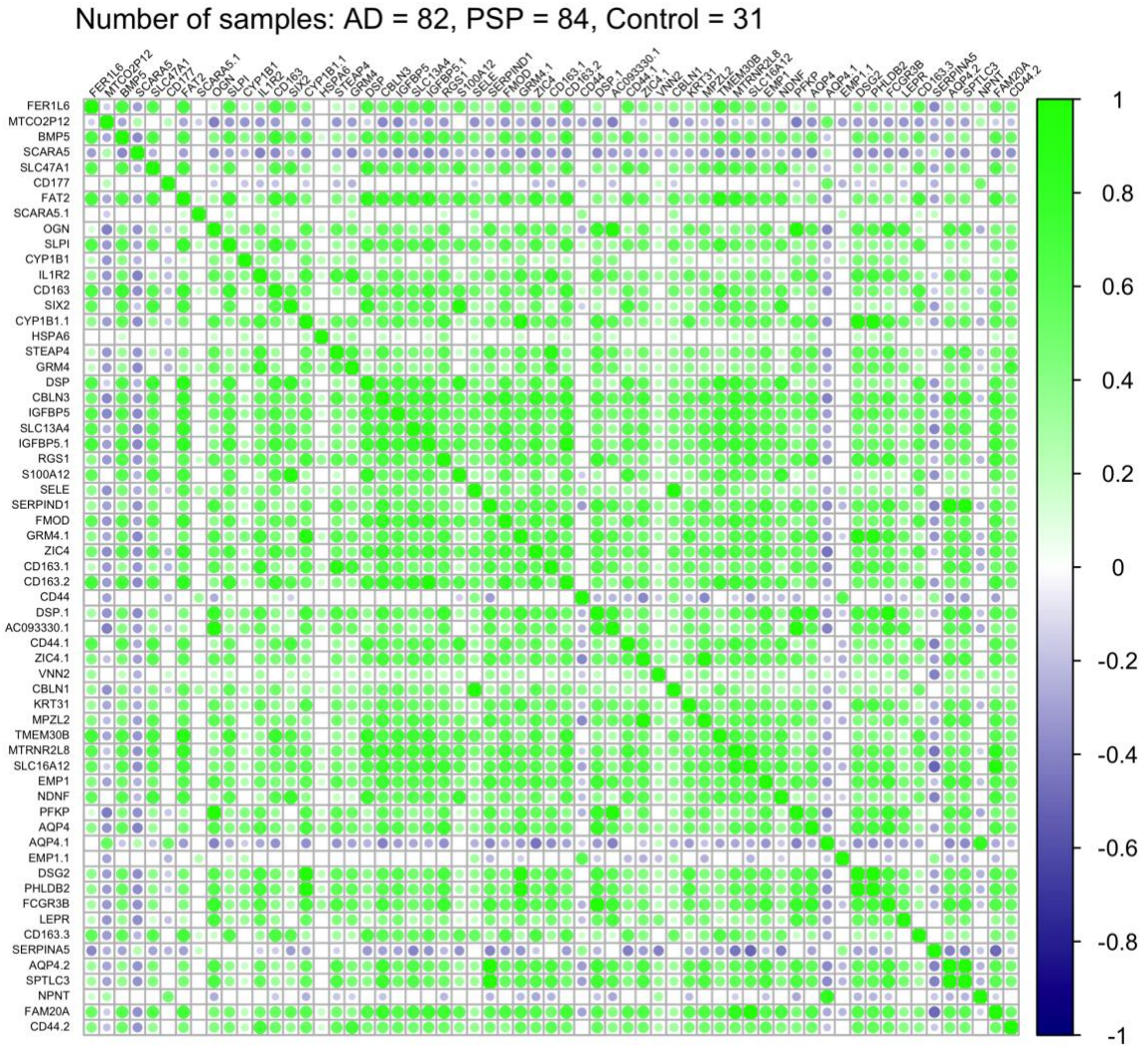


Figure 16.2: Correlation plots of genes involved in AD with genes of all samples

2.4.3: Correlation analysis of proteins and genes involved in AD with proteins and genes in all samples:

We performed correlation between 114 AD proteins and 61 AD genes obtained from AD vs. Control and looked for them in the complete data consisting of AD, Control, and PSP samples.

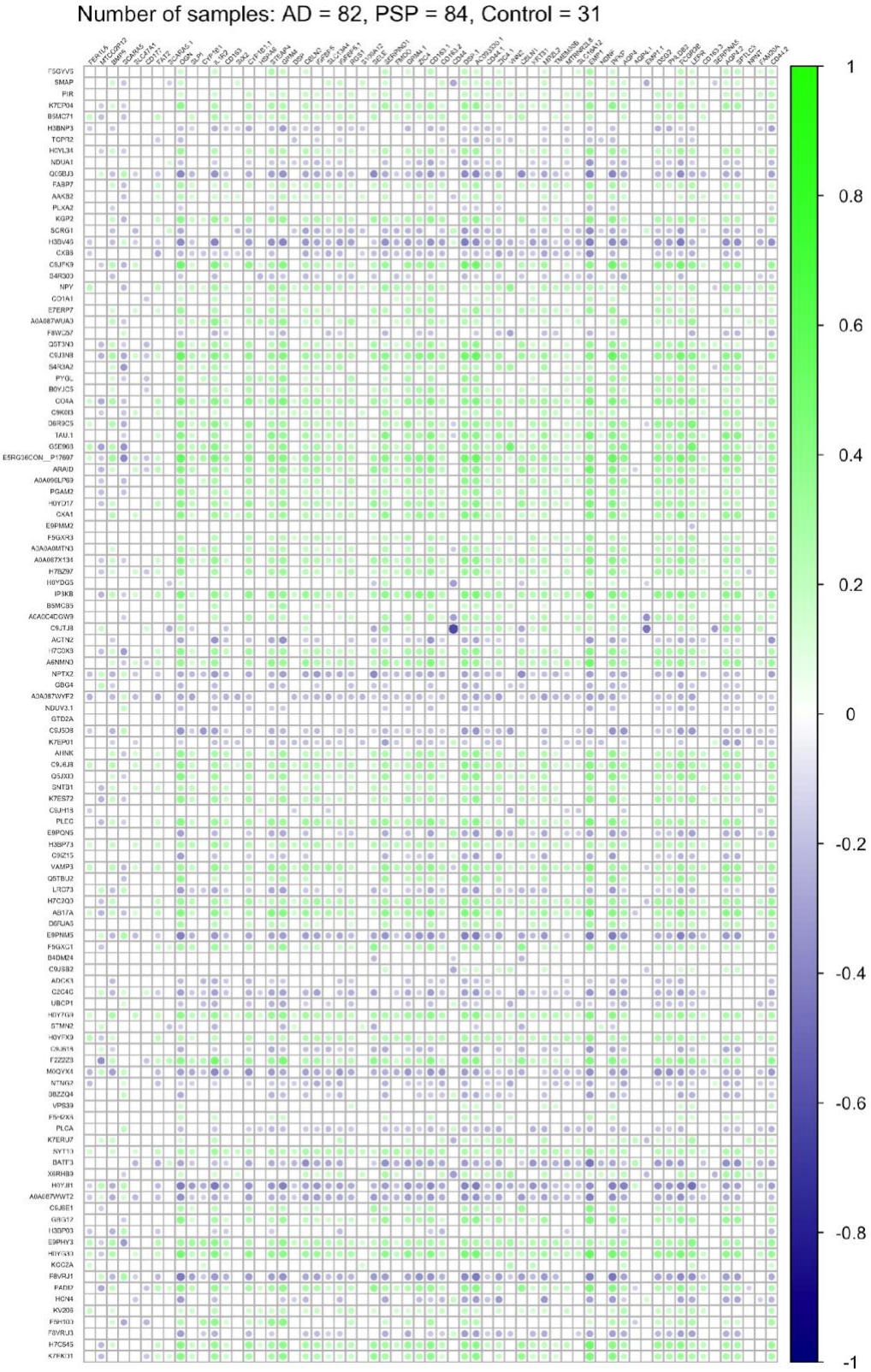

Number of samples: AD = 82, PSP = 84, Control = 31
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 16.2: Correlation plots of proteins and genes involved in AD with all proteins genes of all samples.

Chapter 3: Results and Discussion

3.1: Enrichment analysis:

As mentioned in the methods section, an enrichment analysis consisting of significant DEPs was performed. To see which proteins are functionally related, we performed a functional enrichment analysis for the pairwise comparisons AD vs Control consisting of ARCHS4 Co-expression, KEGG pathways, GO annotations and PPI hub proteins. All of these proteins are seen dysregulated in AD according to our analysis and to a large extent are supported by rich literature evidence. However, cross validation with lab experiments is essential.

3.1.1: KEGG pathways:

We found 42 unique genes involved in 83 pathways. The top genes are involved in ECM-receptor interaction, Calcium signaling pathway, Glucagon signaling pathway, and Alzheimer's Disease pathway.

Table 5.1: KEGG pathways of proteins in AD

Term	Genes
ECM-receptor interaction	SPP1;ITGA6;CD44
Calcium signaling pathway	P2RX7;ATP2B4;PHKA1;PLCD3
Glucagon signaling pathway	PGAM2;PHKA1;PYGL
Alzheimer disease	ADAM10;MAPT;APOE

3.1.2: Gene Ontology:

BP: We found 834 processes shared by 153 unique proteins. The top genes in GO terms involved in biological process like triglyceride catabolic process (GO:0019433) were FABP3; FABP7; APOE, the proteins involved in cellular response to growth factor stimulus (GO:0071363) were ANXA1; STMN2; HSPB1; MAPT; CD44, and extracellular matrix organization (GO:0030198) were FGA; HTRA1; SPP1; ADAM10; ITGA6; CD44respectively.

Table 5.2: Gene Ontology Biological Process pathways of proteins in AD

Term	Genes
triglyceride catabolic process (GO:0019433)	FABP3; FABP7; APOE
cellular response to growth factor stimulus (GO:0071363)	ANXA1; STMN2; HSPB1; MAPT; CD44
extracellular matrix organization (GO:0030198)	FGA; HTRA1; SPP1; ADAM10; ITGA6; CD44

MF: We found 134 molecular functions shared by 51 unique proteins. The top genes in GO terms involved in molecular function like protein homodimerization activity (GO:0042803) were P2RX7;GSTM3;HSPB6;QPRT;HSPB1;ADAM10;PYGL;MAPT;APOE, the proteins involved in intermediate filament binding (GO:0019215) were SYNM;VIM, cadherin binding (GO:0045296) were ANXA1;BAG3;ITGA6;CAPG;PLEC, and calmodulin-dependent protein kinase activity (GO:0004683) were CAMKV;PHKA1 respectively.

Table 5.3: Gene Ontology Molecular Function pathways of proteins in AD

Term	Genes
protein homodimerization activity (GO:0042803)	P2RX7;GSTM3;HSPB6;QPRT;HSPB1;ADAM10;PYGL;MAPT;APOE
intermediate filament binding (GO:0019215)	SYNM;VIM
cadherin binding (GO:0045296)	ANXA1;BAG3;ITGA6;CAPG;PLEC
calmodulin-dependent protein kinase activity (GO:0004683)	CAMKV;PHKA1

CC: We found that 93 cellular components were shared by 40 unique proteins. Proteins involved in Focal adhesion (GO:0005925) were ANXA1; CSRP1; FHL1; HSPB1; ADAM10; ITGA6; VIM; CD44;CD99;SNTB1;PLEC, proteins in the intermediate filament cytoskeleton (GO:0045111) were SYNM;VIM;PLEC, proteins involved in polymeric cytoskeletal fiber (GO:0099513) were SYNM;ANXA1;MAPT;VIM and the proteins involved in intermediate filament (GO:0005882) were SYNM;VIM respectively.

Table 5.4: Gene Ontology Cellular Component pathways of proteins in AD

Term	Genes
focal adhesion (GO:0005925)	ANXA1;CSRP1;FHL1;HSPB1;ADAM10;ITGA6;VIM;CD44;CD99;SNTB1;PLEC
intermediate filament cytoskeleton (GO:0045111)	SYNM;VIM;PLEC
polymeric cytoskeletal fiber (GO:0099513)	SYNM;ANXA1;MAPT;VIM
intermediate filament (GO:0005882)	SYNM;VIM

3.1.3: Protein-Protein Interactions and hub proteins:

Below is the snapshot of protein-protein interactions generated by STRING DB which consider the functional context of proteins based on the pathways and processes taken by the protein sets. For example, the protein ABCC8 in the figure below (figure 11.5 from methods section), is functionally associated with other proteins that have the function - ATP-binding cassette sub-family C member 8; Subunit of the beta-cell ATP-sensitive potassium channel (KATP). Regulator of ATP-sensitive K(+) channels and insulin release; ATP binding cassette subfamily C. The protein ABL1 is functionally associated with proteins that have function as Tyrosine-protein kinase ABL1; Non-receptor tyrosine-protein kinase that plays a role in many key processes linked to cell growth and survival such as cytoskeleton remodeling in response to extracellular stimuli, cell motility and adhesion, receptor endocytosis, autophagy, DNA damage response and apoptosis. Coordinates actin remodeling through tyrosine phosphorylation of proteins controlling cytoskeleton dynamics like WASF3 (involved in branch formation); ANXA1 (involved in membrane anchoring); DBN1, DBNL, CTTN, RAPH1 and ENAH (involved in signaling); or MAPT and PXN (microtubule-binding [...] and ACTA1 is functionally associated with proteins that have function as Actin, alpha skeletal muscle; Actins are highly conserved proteins that are involved in various types of cell motility and are ubiquitously expressed in all eukaryotic cells. We found 156 proteins that act as hub proteins and are associated with other proteins functionally as predicted by STRING DB. (The figure is same as mentioned in methods section figure 11.5, shown here as a reference.)

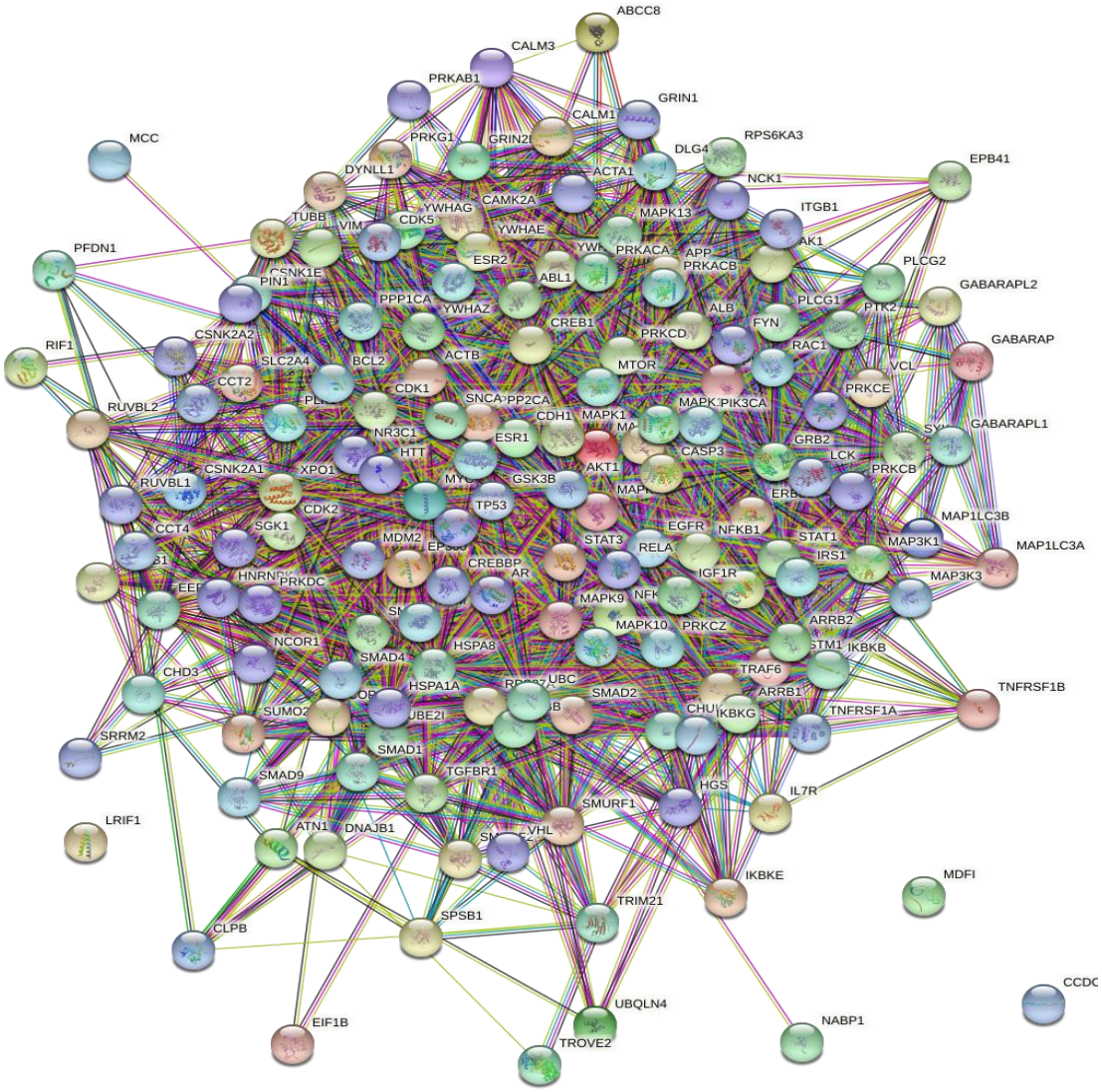


Figure 11.5: Protein-Protein Interactions (PPIs) in AD vs. C from methods section

3.2: Correlation analysis:

When we performed a correlation analysis of proteins involved in AD with proteins in all samples, we found 12,996 protein-protein pairs. We then filtered the correlations at p-value = 0.00001 and got 2520 significant correlations. The mean correlation coefficient was 0.54. We further increased the stringency of the p-value and obtained 192 very significant correlations (p-value <= 0.00000000000005). Mean correlation coefficient = 0.52. (A snapshot of a few is attached below as a table (**Table 5.1**). The whole list is attached as a supplementary table). We find that most proteins form very significant protein-protein pairs.

Table 6.1: Correlations between proteins involved in AD

ProteinID 1	Protein Name	Gene Name	ProteinID 2	Protein Name2	Gene Name 2	Correlation_Coefficient	P_Value
FABP7	Fatty acid-binding protein, brain	FABP7 BLBP, FABPB, MRG	C9J3N8	Heat shock protein beta-1	HSPB1	0.541015663	2.22E-16
TAU.1	Microtubule-associated protein tau	MAPT MAPTL, MTBT1, TAU	AHNK	Neuroblast differentiation-associated protein AHNAK	AHNAK PM227	0.541330576	2.22E-16
H7C0X8	Macrophage-capping protein	CAPG	PLEC	Plectin	PLEC PLEC1	0.540129387	2.22E-16
C9J3N8	Heat shock protein beta-1	HSPB1	VAMP3	Vesicle-associated membrane protein	VAMP3 SYB3	0.541696	2.22E-16
CXA1	Gap junction alpha-1 protein	GJA1 GJAL	E9PNM5	Synaptotagmin-12	SYT12	-0.541975635	2.22E-16
PLEC	Plectin	PLEC PLEC1	H0Y7G9	Serine protease HTRA1	HTRA1	0.540607972	2.22E-16
B0YJC5	Vimentin	VIM	F8VRJ1	Rabphilin-3A	RPH3A	-0.542964494	2.22E-16
H0YL34	Synemin	SYNM	PADI2	Protein-arginine deiminase type-2	PADI2 KIAA0994, PAD2, PDI2	0.540138421	2.22E-16
E9PNM5	Synaptotagmin-12	SYT12	K7EKD1	Glial fibrillary acidic protein	GFAP	-0.541723337	2.22E-16

FABP7	Fatty acid-binding protein, brain		B0YJC5	Vimentin	VIM	0.536989204	4.44E-16
B5MC71	Disintegrin and metalloproteinase domain-containing protein 10	ADAM10	A0A087X134	Cysteine and glycine-rich protein 1	CSRP1	0.534652075	4.44E-16
H0YL34	Synemin	SYNM	IP3KB	Inositol-trisphosphate 3-kinase B	ITPKB	0.536455906	4.44E-16
NPTX2	Neuronal pentraxin-2	NPTX2	PLEC	Plectin	PLEC PLEC1	-0.536346138	4.44E-16
Q5JXI0	Four and a half LIM domains protein 1	FHL1	VAMP3	Vesicle-associated membrane protein	VAMP3 SYB3	0.538409103	4.44E-16
FABP7	Fatty acid-binding protein, brain	FABP7 BLBP, FABPB, MRG	Q5TBU2	Adipogenesis regulatory factor	ADIRF	0.537125526	4.44E-16
CO4A	Complement C4-A	C4A CO4, CPAMD2	E9PNM5	Synaptotagmin-12	ADAM10	-0.534962468	4.44E-16
H0YD17	CD44 Antigen	CD44	C2C4C	C2 calcium-dependent domain-containing protein 4C	C2CD4C FAM148C, KIAA1957, NLF3	-0.537908635	4.44E-16
C9J6J8	Sequestosome-1	SQSTM1	A0A087WWT2	Neuritin	NRN1	-0.535671513	4.44E-16
E5RG36CON__P17697	No information	No Information	ARAID	All-trans retinoic acid-induced differentiation factor	ATRAID APR3, C2orf28, HSPC013, UNQ214/PRO240	0.533656744	6.66E-16
C9IZ15	TSC22 domain family protein 1	TSC22D1	GBG12	Guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma 12	GNG12	-0.534341712	6.66E-16

Similarly, we performed a correlation analysis of genes involved in AD with genes in all samples, we found 3,721 gene-gene pairs. We then filtered the correlations at p-value = 0.00001 and got 2266 significant correlations. The mean correlation coefficient was 0.54. We then further increased the

stringency of the p-value and obtained 208 very significant correlations (p value <= 0.0000000000005). Mean correlation coefficient = 0.52. (A snapshot of a few is attached below in figure. The whole list is attached).

Table 6.2: Correlations between genes involved in AD

Gene ID1	Protein Name	Gene Name	GeneID2	Protein Name2	Gene Name2	Correlation Coefficient	P-Value
CD44.1	CD44 antigen	CD44 LHR, MDU2, MDU3, MIC4	EMP1	**Epithelial membrane protein 1**	**EMP1** B4B, TMP	0.540784934	2.22E-16
CBLN3	**Cerebellin-3**	CBLN3 UNQ755/PRO1486	AC093330.1	Zinc finger protein 236	ZNF236	0.542828576	2.22E-16
SPTLC3	Serine palmitoyltransferase 3	SPTLC3 C20orf38, SPTLC2L	FMOD	**Fibromodulin**	FMOD FM, SLRR2E	0.54162311	2.22E-16
PHLDB2	Pleckstrin homology-like domain family B member 2	PHLDB2 LL5B	LEPR	Leptin receptor	LEPR DB, OBR	0.542814584	2.22E-16
LEPR	**Leptin receptor**	LEPR DB, OBR	DSG2	**Desmoglein-2**	DSG2 CDHF5	0.542721976	2.22E-16
CYP1B1.1	**Cytochrome P450 1B1**	CYP1B1	IGFBP5.1	Insulin-like growth factor-binding protein 5	**IGFBP5** IBP5	0.534479782	4.44E-16
GRM4	**Metabotropic glutamate receptor 4**	GRM4 GPRC1D, MGLUR4	RGS1	**Regulator of G-protein signaling 1**	RGS1 1R20, BL34, IER1	0.53933669	4.44E-16
CBLN3	Cerebellin-3	CBLN3 UNQ755/PRO1486	S100A12	**Protein S100-A12**	S100A12	0.536169759	4.44E-16
BMP5	Bone morphogenetic protein 5	**BMP5**	GRM4.1	Metabotropic glutamate receptor 4	GRM4 GPRC1D, MGLUR4	0.536069553	4.44E-16
OGN	**Mimecan**	OGN OIF, SLRR3A	ZIC4	**Zinc finger protein ZIC 4**	ZIC4	0.537741446	4.44E-16
SLC13A4	**Solute carrier family 13 member 4**	SLC13A4 SUT1	KRT31	**Keratin, type I cuticular Ha1**	KRT31 HHA1, HKA1, KRTHA1	0.536924487	4.44E-16
DSP.1	**Desmoplakin**	DSP	MPZL2	**Myelin protein zero-like protein 2**	MPZL2 EVA, EVA1, UNQ606/PRO1192	0.536149854	4.44E-16
IL1R2	**Interleukin-1 receptor type 2**	IL1R2 IL1RB	EMP1	Epithelial membrane protein 1	EMP1 B4B, TMP	0.536868805	4.44E-16
ZIC4.1	**Zinc finger protein ZIC 4**	ZIC4	FCGR3B	Low affinity immunoglobulin gamma Fc region receptor III-B	**FCGR3B** CD16B, FCG3, FCGR3, IGFR3	0.539361291	4.44E-16
SELE	**E-selectin**	SELE ELAM1	CD163.3	Scavenger receptor cysteine-rich type 1 protein M130	CD163 M130	0.535336793	4.44E-16
RGS1	Regulator of G-protein signaling 1	RGS1 1R20, BL34, IER1	GRM4	Metabotropic glutamate receptor 4	GRM4 GPRC1D, MGLUR4	0.53933669	4.44E-16
S100A12	Protein S100-A12	S100A12	CBLN3	Cerebellin-3	CBLN3 UNQ755/PRO1486	0.536169759	4.44E-16
KRT31	**Keratin, type I cuticular Ha1**	KRT31 HHA1, HKA1, KRTHA1	SLC13A4	Solute carrier family 13 member 4	SLC13A4 SUT1	0.536924487	4.44E-16

We performed a correlation analysis of proteins and genes involved in AD with proteins and genes in all samples, we found 6954 protein-gene pairs that were correlated. We then filtered the correlations (p-value = 0.0001) and got 395 significant correlations. The correlation coefficient was a high value of 0.53. (A snapshot of a few is attached below in figure. The whole list is attached).

Table 6.3: Correlation analysis of proteins and genes involved in AD

ProteinID	Protein Name Gene Name	GeneID	Protein Name	Correlation Coefficient	P-Value
TAU.1	Microtubule-associated protein tau-MAPT MAPTL, MTBT1, TAU	H7BZ97	Integrin alpha-6	0.48589622	4.56E-13
C9JFK9	BAG family molecular chaperone regulator 3-BAG3	G5E968	Chromogranin A (Parathyroid secretory protein 1), isoform CRA_b-CHGA	0.48109284	8.32E-13
E5RG36CON__P17697	No information	A6NMN0	Phosphorylase b kinase regulatory subunit	0.47614383	1.53E-12
C9JFK9	BAG family molecular chaperone regulator 3-BAG3	A6NMN0	Phosphorylase b kinase regulatory subunit	0.47108802	2.82E-12
CO4A	Complement C4-A-C4A CO4, CPAMD2	H7BZ97	Integrin alpha-6	0.4586358	1.22E-11
E5RG36CON__P17697	No information	H7BZ97	Integrin alpha-6	0.45235458	2.50E-11
D6R9C5	Osteopontin-SPP1	NPTX2	Neuronal pentraxin-2	0.45120781	2.85E-11
H0YD17	CD44 antigen-CD44	H7BZ97	Integrin alpha-6	0.44936399	3.50E-11
CO4A	Complement C4-A-C4A CO4, CPAMD2	IP3KB	Inositol-trisphosphate 3-kinase B	0.44391583	6.41E-11
CO4A	Complement C4-A-C4A CO4, CPAMD2	E5RG36CON__P17697	No infprmation	0.43889773	1.11E-10
ARAID	All-trans retinoic acid-induced differentiation factor-ATRAID APR3, C2orf28, HSPC013, UNQ214/PRO240	IP3KB	Inositol-trisphosphate 3-kinase B	0.43510295	1.67E-10
H3BV46	BAI1-associated protein 3-BAI1AP3	G5E968	Chromogranin A (Parathyroid secretory protein 1), isoform CRA_b-CHGA	-0.428894	3.21E-10
H3BV46	BAI1-associated protein 3-BAI1AP3	A6NMN0	Phosphorylase b kinase regulatory subunit	-0.4287637	3.26E-10
C9J3N8	Heat shock protein beta-1-HSPB1	H7BZ97	Integrin alpha-6	0.42790063	3.57E-10
CO4A	Complement C4-A-C4A CO4, CPAMD2	C9K0I3	Bifunctional peptidase and (3S)-lysyl hydroxylase JMJD7	0.42766861	3.65E-10

We then wanted to look at the patterns of correlations between proteins and genes. So, we first sorted the proteins alphabetically with their corresponding correlation coefficients and p-values. It was fascinating to observe that there were at least 30 protein-gene pairs in which a single protein was found correlating with multiple genes. An instance is shown in the table below. The protein name for tau is

Microtubule-associated protein tau and can be found in biological databases as MAPT, MAPT1, MTBT1, and TAU. As already mentioned in the introduction section, hyperphosphorylation of tau protein is a hallmark of neurodegeneration in AD. Here in the table 5.4 below, we can see that tau protein shows significant correlations with at least 14 genes namely, H7BZ97 (Integrins – delay the loss of neurons in AD), GTD2A (several transcription factors are involved in AD regulatory networks), NDUV3 (contributes to alterations of oxidative metabolism in AD), IP3KB (increased in human AD), E5RG36CON_P17697 (no information, could be novel protein/gene which requires evidence at molecular level), A6NMN0 (gene name: PHKA1 differentially expressed in AD mice), GSE968 (gene name: CHGA is downregulated in AD), A0A096LP69 (CD99 antigen plays a key role in AD inflammation), E9PMM2 (UBTF: upstream nuclear transcription factor is a nuclear chaperone whose expression is decreased in AD), AHNK (literature suggests a strong connect between APOE gene and AHNK), NPTX2 (reduced expression in AD), H7C0X8 (macrophage capping protein), ACTN2 (highly enriched in AD) and B5MCB5 (CRABP1: contributes to AD neurogenesis; is upregulated in AD) [31-42]

Table 6.4: Correlation analysis depicting association of a single protein with multiple genes involved in AD

Protein	Protein Name + Gene Name	Gene	Protein Name Gene Name	Correlation Coefficient	P-Value
TAU.1	Microtubule-associated protein tau-MAPT MAPTL, MTBT1, TAU	H7BZ97	Integrin alpha-6	0.48589622	4.56E-13
TAU.1		GTD2A	General transcription factor II-I repeat domain-containing protein 2A, GTF2I repeat domain-containing protein 2A (Transcription factor GTF2IRD2-alpha)-GTF2IRD2 GTF2IRD2A	0.42084887	7.38E-10
TAU.1		NDUV3.1	NADH dehydrogenase [ubiquinone] flavoprotein 3- NDUFV3	0.42050355	7.65E-10
TAU.1		IP3KB	Inositol-trisphosphate 3-kinase B	0.41674976	1.12E-09
TAU.1		E5RG36CON__P17697	no information	0.40445714	3.76E-09
TAU.1		A6NMN0	Phosphorylase b kinase regulatory subunit	0.38708047	1.92E-08
TAU.1		G5E968	Chromogranin A (Parathyroid secretory protein 1), isoform CRA_b-CHGA	0.38519377	2.28E-08
TAU.1		A0A096LP69	CD99 antigen-CD99	0.34653516	6.09E-07
TAU.1		E9PMM2	Nucleolar transcription factor 1-UBTF	0.34325619	7.89E-07
TAU.1		AHNK	Neuroblast differentiation-associated protein-AHNAK PM227	0.33054421	2.10E-06
TAU.1		NPTX2	Neuronal pentraxin-2	0.31213234	8.00E-06
TAU.1		H7C0X8	Macrophage-capping protein-CAPG	0.27227525	0.00010855
TAU.1		ACTN2	Alpha-actinin-2	0.27171175	0.00011232
TAU.1		B5MCB5	Cellular retinoic acid-binding protein 1-CRABP1	0.25603612	0.00028179

Similar to above correlations where a single protein is found to be associated with multiple genes, it was also fascinating to observe that a single gene is forming pairs with multiple proteins, which was obtained by sorting genes alphabetically along with their corresponding correlation coefficients and p-values. We found more than 32 such correlations, an instance can be seen in the table below, most of which are dysregulated in AD.

Table 6.5: Correlation analysis depicting association of a single gene with multiple proteins involved in AD

Protein	Protein Name + Gene Name	Gene	Protein Name + Gene Name2	Correlation Coefficient	P-Value
C9J3N8	Heat shock protein beta-1-HSPB1	A6NMN0	Phosphorylase b kinase regulatory subunit-PHKA1	0.489925559	2.74E-13
E5RG36CON__P17697	No information	A6NMN0		0.476143835	1.53E-12
C9JFK9	BAG family molecular chaperone regulator 3-BAG3	A6NMN0		0.471088024	2.82E-12
H3BV46	BAI1-associated protein 3-BAI1AP3	A6NMN0		-0.428763713	3.26E-10
CXA1	Gap junction alpha-1 protein-GJA1 GJAL	A6NMN0		0.41843321	9.43E-10
IP3KB	Inositol-trisphosphate 3-kinase B	A6NMN0		0.411710575	1.85E-09
TAU.1	Microtubule-associated protein tau-MAPT MAPTL, MTBT1, TAU	A6NMN0		0.38708047	1.92E-08
CO4A	Complement C4-A-C4A CO4, CPAMD2	A6NMN0		0.381895129	3.07E-08
A0A096LP69	CD99 antigen-CD99	A6NMN0		0.378411298	4.19E-08
A0A087X134	Cysteine and glycine-rich protein 1-CSRP1	A6NMN0		0.376748431	4.85E-08
Q05BJ3	Neurosecretory protein VGF-VGF	A6NMN0		-0.362933166	1.59E-07
K7EP04	Heat shock protein beta-6-HSPB6	A6NMN0		0.353055713	3.61E-07
PYGL	Glycogen phosphorylase, liver form	A6NMN0		0.352518709	3.77E-07
D6R9C5	Osteopontin-SPP1	A6NMN0		0.343557242	7.71E-07
H0YD17	CD44 antigen-CD44	A6NMN0		0.342570037	8.33E-07
ARAID	All-trans retinoic acid-induced differentiation factor-ATRAID APR3, C2orf28, HSPC013, UNQ214/PRO240	A6NMN0		0.341700676	8.91E-07
H7BZ97	Integrin alpha-6	A6NMN0		0.340990472	9.42E-07
H0YL34	Synemin-SYNM	A6NMN0		0.333749153	1.65E-06
Q5T3N0	Annexin-ANXA1	A6NMN0		0.333194421	1.72E-06
A0A0A0MTN3	Glutathione S-transferase-GSTM3	A6NMN0		0.323039782	3.66E-06
H7C0X8	Macrophage-capping protein-CAPG	A6NMN0		0.322440593	3.82E-06
A0A0C4DGW9	Serpin I2-SERPINI2	A6NMN0		0.308083085	1.06E-05
SCRG1	Scrapie-responsive protein 1-SCRG1 UNQ390/PRO725	A6NMN0		-0.308082249	1.06E-05
KGP2	cGMP-dependent protein kinase 2-PRKG2, PRKGR2	A6NMN0		0.306619938	1.17E-05
B0YJC5	Vimentin-VIM	A6NMN0		0.301276314	1.69E-05
G5E968	Chromogranin A (Parathyroid secretory protein 1), isoform CRA_b-CHGA	A6NMN0		0.295936907	2.42E-05
A0A087WUA0	Fibrinopeptide A-FGA	A6NMN0		0.27012561	0.000123598
NDUA1	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 1	A6NMN0		-0.269790811	0.00012611
PGAM2	Phosphoglycerate mutase 2-PGAM2, PGAMM	A6NMN0		0.268580719	0.000135591
S4R3A2	Fatty acid-binding protein, heart-FABP3	A6NMN0		0.268317615	0.000137739

Below is the interpretation of pairing between proteins and genes**:**

Table 6.6: Biological interpretations of the correlations between proteins and genes with r >=0.45. Correlation analysis depicting association of proteins with genes involved in AD to see the direction of association patterns

Protein	Gene	Correlation Coeff.	P-val
C9J3N8	E5RG36CON__P17697	0.5348593	4.44E-16
E5RG36CON__P17697	IP3KB	0.53832742	4.44E-16
C9JFK9	IP3KB	0.50479345	3.91E-14
G5E968	NPTX2	0.49616811	1.22E-13
C9JFK9	E5RG36CON__P17697	0.49506646	1.41E-13
G5E968	PGAM2	0.49435282	1.55E-13
C9J3N8	G5E968	0.49295656	1.86E-13
C9J3N8	A6NMN0	0.48992556	2.74E-13
ARAID	H7BZ97	0.48963984	2.84E-13
TAU.1	H7BZ97	0.48589622	4.56E-13
C9JFK9	G5E968	0.48109284	8.32E-13
E5RG36CON__P17697	A6NMN0	0.47614383	1.53E-12
C9JFK9	A6NMN0	0.47108802	2.82E-12
CO4A	H7BZ97	0.4586358	1.22E-11
E5RG36CON__P17697	H7BZ97	0.45235458	2.50E-11
D6R9C5	NPTX2	0.45120781	2.85E-11
Protein	**Gene**	**Corelation Coeff.**	**P-value**
H3BV46	G5E968	-0.428894	3.21E-10
H3BV46	A6NMN0	-0.42876371	3.26E-10
H3BV46	H7BZ97	-0.41958821	8.39E-10
Q05BJ3	H7BZ97	-0.4119914	1.80E-09

Correlations between C9J3N8, IP3KB, C9JFK9, A6NMN0, H7BZ97, and E5RG36CON__P17697:

C9J3N8 is a heat shock protein (HSP Beta 1). Heat shock proteins (HSPs) like Hsp60, and Hsp70 that play the role of molecular chaperones in the cell have a range of functions to maintain cellular homeostasis when they face stress. They are found associated with stopping APP/Aβ protein folding, in that they can prevent the aggregation of unfolded proteins in AD [43]. Our analysis shows that

C9J3N8 is highly correlated (r = 0.53) with a gene of unknown information (E5RG36CON__P17697) in the biological database (Ensembl). While we could find all the proteins and genes are annotated in the biological databases like UniProt and Ensembl, there was a protein by the ID 'E5RG36CON__P17697' which was found significantly correlated at both gene and protein levels, but we found no information about it. So, this could be our novel gene/protein. However, it can be further confirmed with immunoprecipitation/immunofluorescence protocols. The study of these high correlations between HSPs and AD is a significant area of research. mRNA levels of ITPKB gene (protein name Inositol-trisphosphate 3-kinase B and gene ID IP3KB) are significantly increased in AD [44]. This gene is seen forming protein - gene pairs with protein E5RG36CON__P17697 and protein C9JFK9 having very high correlations r >= 0.50. C9JFK9 (BAG family molecular chaperone regulator 3) also belongs to the family of molecular chaperones. BAG3 facilitates clearing off of tau protein, which, due to its hyperphosphorylation, causes neurodegeneration and aggressiveness of AD [45]. From the above information, and based on the correlations, we observe that E5RG36CON__P17697 has a positive relationship with C9J3N8, IP3KB, C9JFK9. This gives us an insight that E5RG36CON__P17697 might also be a protein/gene which might be associated with a function that opposes the progress of AD. A6NMN0 is a subunit of PHKG2 (Phosphorylase b kinase gamma catalytic chain) and mediates neuronal regulation of glycogen breakdown by phosphorylating and thereby activating glycogen phosphorylase. The catalysis of the reaction: ATP + tau-protein = ADP + O-phospho-tau-protein. There is a tight coupling between glycogen breakdown and neuronal development. Studies show that glucose levels are reduced in AD [46]. While we discussed above that E5RG36CON__P17697 is associated with molecular chaperones, which aid in stopping the advancement of AD, we also observe that E5RG36CON__P17697 is highly correlated (r = 0.47)

with A6NMN0 which mediates glucose metabolism. Aβ interacts with integrin subunits (H7BZ97/ITGA6) in the early onset of AD [47,48].

Correlations between C9JFK9 and IP3KB and association with E5RG36CON__P17697:

C9JFK9 is a molecular chaperone that interacts with ITPKB gene whose mRNA levels are increased in AD. E5RG36CON__P17697 positively correlated with ITPKB and C9JFK9.

We can see from the associations that these proteins and genes could probably be very well associated with AD and could be significant research areas for their therapeutics.

Correlations of G5E968 with NPTX2, PGAM2, C9J3N8, C9JFK9, and H3BV46:

In AD, about 30% of beta-amyloid plaques co-labeled with proteins chromogranin A (Gene ID G5E968). A study shows that Chromogranin A might be a mediator amongst neuronal, glial, and inflammatory mechanisms found in AD [49].

Neuronal pentraxin-2 (gene ID NPTX2) levels are reduced in AD leading to cognitive dysfunction [50].

PGAM2 is a hub protein that is enriched in AD-related pathways [51]. A study by Chen, Lu et al., 2019 shows that BAI1-associated protein 3 (Gene ID H3BV46) is downregulated in AD. We observe from our analysis that G5E968 is positively correlated with NPTX2, PGAM2, C9J3N8, C9JFK9 (r = 0.49) and negatively correlated with H3BV46 (r = -0.42).

Correlation of H7BZ97 with ARAID, TAU.1, CO4A, E5RG36CON__P17697 and H3BV46, Q05BJ3:

H7BZ97 is also an integrin (protein name – Integrin alpha 6 - ITGA6). As discussed above, Aβ protein that forms amyloid plaques in AD, interacts with integrin subunits. We couldn't find enough information about All-trans retinoic acid-induced differentiation factor (gene name – ATRAID) and its

association with AD. CO4A is associated with neuroinflammation in AD [52]. TAU.1 (MAPT1) is highly hyperphosphorylated in AD [53]. As already discussed above, BAI1-associated protein 3 (H3BV46) is downregulated in AD [54]. Peptides derived from VGF (Q05BJ3 (Protein name-Neurosecretory protein VGF) are reduced in cerebrospinal fluid of AD patients which helps against advancement of AD [55,56]. From our analysis, we see that H7BZ97 is positively correlated with ARAID (r = 0.48), TAU.1(r = 0.48), CO4A (r = 0.45), and E5RG36CON__P17697 (r = 0.45) but is negatively correlated with H3BV46 (r = -0.41) and Q05BJ3 (r = -0.41). However, all of the proteins and genes are associated with AD.

Chapter 4: Conclusion

Genes are the basic unit of hereditary material in any living organism. Proteins are complex molecules that play critical roles in an organism. Most genes contain the information required to make proteins but the process of formation of proteins from genes is a complex journey within the cell which involves complex gene regulation. So, it is difficult to understand the relationship between them. There are also factors like transcription factors, RNA binding proteins (RBPs) and regulatory factors should be studied if we were to study the interactions between genes and proteins. A study by Koussounadis et al., 2015 shows that significantly higher correlations occur between mRNA and protein for genes with differentially expressed mRNA. Zaborowski and Walther, 2020 in their paper mentioned that while transcription factors (TFs) are known to regulate the expression of their target genes (TGs), only a weak correlation of expression between TFs and their TGs has generally been observed. So, since ours is a data driven approach and validation via immunoprecipitation is further needed to understand the interactions or interplay between genes and proteins, and while it's known that correlation doesn't imply causation, we just wanted to see what correlations existed between proteins and genes without a focus on the actual interactions. From our analysis, we have seen that TAU protein, APOE genes, and amyloid-beta (Aβ) proteins are significantly dysregulated in AD and play a significant role in the progression of the disease in accordance with literature and past research. We also observed that they are seen being in an association with a number of other proteins and genes that further advance the disease and are also seen pairing with proteins and genes that act as molecular chaperones which is a very interesting area of research. It was also fascinating to see the gene/protein E5RG36CON__P17697 is associated in different ways with a set of proteins and genes. We can thus conclude that there could be a probability that these numerous genes and proteins involved in the disease progression pathways

and processes as well as those that act in opposition to neurodegeneration could probably either be biomarkers that can detect the disease at an early stage and/or be the targets for therapeutics that are disease-modifying. To be able to understand what actually the mechanism at the molecular level is, further research at molecular level like co-immunoprecipitation is strongly needed.

Chapter 5: Challenges in the current work

Whereas analysis of individual datasets is not very robust for the derivation of biological insights, large datasets obtained from high throughput technologies have revolutionized biology research. However, it is challenging to work with these huge datasets. The three most important areas of the challenge with -omics datasets are integration, interpretation, and extracting biological insights from the data (Misra et al., 2018). Before the integration of the data, each data set has to be filtered, cleaned, imputed, normalized, and scaled in such a way that the biological insights are reached. Correcting the proteomics dataset for batch effects was the most time taking and challenging step. Secondly, there is a large variation of sample sizes and abundance quantities which had to be taken care of. Thirdly, when integrating the DEPs and DEGs, the most challenging part was to look for them in the raw datasets based on the experimental conditions. We wanted to look for the proteins and genes differentially expressed in the pairwise comparisons (AD vs. C) in the complete data consisting of AD, PSP, and Control samples. After having done this step, given thousands of significant correlated proteins and genes, it was a challenge to search the literature to understand the biological insights that aligned with our analysis to at least some extent. So, we searched the literature for 20 most significant ones.

Chapter 6: Future work

Studies show that a lot of research is being done in various ways to study AD. With the evolution of next-generation sequencing technologies and mass spectrometry technologies, it is becoming possible to use artificial intelligence in healthcare. It is thus getting more interesting to study disease causes and correlations. In the current study, we procured proteomics and RNA seq data to perform analysis. Initially, as we began, we performed a pairwise differential expression of AD vs. C, PSP vs. C and AD vs. PSP for both the datasets. In the correlation analysis, we focused only on AD proteins and genes. One great focus scope is definitely to look at the correlations with respect to PSP which is not done in this project. Future work could probably also include the integration and correlation analysis of more datasets that would facilitate in studying the aspects of epigenetics (DNA methylation, histone modifications, chromatin accessibility transcription binding sites), SNPs, CNVs, SAVs, loss of heterozygosity, finding rare variants, etc. to get a complete picture of the intricate molecular processes involved in the disease. Another good way could be making use of univariate, multivariate, and predictive analytics (machine learning and deep learning models) that can help predict the disease in its early stages.

REFERENCES

1. Mirra S. S. (1997). The CERAD neuropathology protocol and consensus recommendations for the postmortem diagnosis of Alzheimer's disease: a commentary. *Neurobiology of aging*, *18*(4 Suppl), S91–S94.

2. Lee, S. J., Lim, H. S., Masliah, E., & Lee, H. J. (2011). Protein aggregate spreading in neurodegenerative diseases: problems and perspectives. *Neuroscience research*, *70*(4), 339–348.

3. Murphy, M. P., & LeVine, H., 3rd (2010). Alzheimer's disease and the amyloid-beta peptide. *Journal of Alzheimer's*

4. Yiannopoulou, K. G., & Papageorgiou, S. G. (2013). Current and future treatments for Alzheimer's disease. *Therapeutic advances in neurological disorders*, *6*(1), 19–33.

5. O'Brien, R. J., & Wong, P. C. (2011). Amyloid precursor protein processing and Alzheimer's disease. *Annual review of neuroscience*, *34*, 185–204.

6. H. R. Morris, G. Gibb, R. Katzenschlager, N. W. Wood, D. P. Hanger, C. Strand, T. Lashley, S. E. Daniel, A. J. Lees, B. H. Anderton, T. Revesz, Pathological, clinical and genetic heterogeneity in progressive supranuclear palsy, *Brain*, Volume 125, Issue 5, May 2002, Pages 969–975.

7. Singhal, P., Verma, S. S., Dudek, S. M., & Ritchie, M. D. (2019). Neural network-based multiomics data integration in Alzheimer's disease. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*.

8. Dragomir, A., Vrahatis, A. G., & Bezerianos, A. (2019). A Network-Based Perspective in Alzheimer's Disease: Current State and an Integrative Framework. *IEEE Journal of Biomedical and Health Informatics, 23*(1), 14-25.

9. Avramouli, A., & Vlamos, P. M. (2017). Integrating Omic Technologies in Alzheimer's Disease. *Advances in Experimental Medicine and Biology GeNeDis 2016*, 177-184.

10. Bonvicini, C., Scassellati, C., Benussi, L., Di Maria, E., Maj, C., Ciani, M., Fostinelli, S., Mega, A., Bocchetta, M., Lanzi, G., Giacopuzzi, E., Ferraboli, S., Pievani, M., Fedi, V., Defanti, C. A., Giliani, S., Alzheimer's Disease Neuroimaging Initiative, Frisoni, G. B., Ghidoni, R., & Gennarelli, M. (2019). Next Generation Sequencing Analysis in Early Onset Dementia Patients. *Journal of Alzheimer's disease : JAD*, *67*(1), 243–256.

11. Cox, J. et al. Journal of Proteome Research, 2011, 10, p. 1794-1805.

12. Varma, A. R., Snowden, J. S., Lloyd, J. J., Talbot, P. R., Mann, D. M., & Neary, D. (1999). Evaluation of the NINCDS-ADRDA criteria in the differentiation of Alzheimer's disease and frontotemporal dementia. *Journal of neurology, neurosurgery, and psychiatry*, *66*(2), 184–188.

13. Purcell, S., et al. Am J Hum Genet, 2007. 81(3): p. 559-75.

14. Price, A.L., et al. Nat Genet, 2006. 38(8): p. 904-9.

15. Wang, L., et al. Bioinformatics, 2012. 28(16): p. 2184–5.

16. Zaharia et al. arXiv:1111.5572, 2011.

17. Dolan, P. J., & Johnson, G. V. (2010). The role of tau kinases in Alzheimer's disease. *Current opinion in drug discovery & development*, *13*(5), 595–603.

18. Soto, C., & Pritzkow, S. (2018). Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nature neuroscience*, *21*(10), 1332–1340.

19. Brenes, A., Hukelmann, J., Bensaddek, D., & Lamond, A. I. (2019). Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. *Molecular & cellular proteomics : MCP*, *18*(10), 1967–1980.

20. Pérez Diz, Ángel & Truebano, Manuela & Skibinski, David. (2009). The consequences of sample pooling in proteomics: An Empirical Study. Electrophoresis. 30. 2967-75. 10.1002/elps.200900210.

21. Johnson, E.C.B., Dammer, E.B., Duong, D.M. *et al.* Deep proteomic network analysis of Alzheimer's disease brain reveals alterations in RNA binding proteins and RNA splicing associated with disease. *Mol Neurodegeneration* **13,** 52 (2018).

22. Ping, L., Duong, D., Yin, L. *et al.* Global quantitative analysis of the human brain proteome in Alzheimer's and Parkinson's Disease. *Sci Data* **5,** 180036 (2018).

23. Molinari, N., Roche, S., Peoc'h, K., Tiers, L., Séveno, M., Hirtz, C., & Lehmann, S. (2018). Sample Pooling and Inflammation Linked to the False Selection of Biomarkers for Neurodegenerative Diseases in Top-Down Proteomics: A Pilot Study. *Frontiers in molecular neuroscience*, *11*, 477.

24. Johnson, E., Dammer, E. B., Duong, D. M., Ping, L., Zhou, M., Yin, L., Higginbotham, L. A., Guajardo, A., White, B., Troncoso, J. C., Thambisetty, M., Montine, T. J., Lee, E. B., Trojanowski, J. Q., Beach, T. G., Reiman, E. M., Haroutunian, V., Wang, M., Schadt, E., Zhang, B., … Seyfried, N. T. (2020). Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nature medicine*, *26*(5), 769–780.

25. Johnson, E.C.B., Dammer, E.B., Duong, D.M. *et al.* Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat Med* **26,** 769–780 (2020).

26. Tommi Välikangas, Tomi Suomi, Laura L Elo, A systematic evaluation of normalization methods in quantitative label-free proteomics, *Briefings in Bioinformatics*, Volume 19, Issue 1, January 2018, Pages 1–11.

27. Jafari, M., & Ansari-Pour, N. (2019). Why, When and How to Adjust Your P Values?. *Cell journal*, *20*(4), 604–607.

28. Eva Brombacher, Ariane Schad, Clemens KreutzbioRxiv 2020.04.17.046227;

29. Chen, E.Y., Tan, C.M., Kou, Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14,** 128 (2013).

30. Gry, M., Rimini, R., Strömberg, S., Asplund, A., Pontén, F., Uhlén, M., & Nilsson, P. (2009). Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC genomics*, *10*, 365.

31. Wu, X., & Reddy, D. S. (2012). Integrins as receptor targets for neurological disorders. *Pharmacology & therapeutics*, *134*(1), 68–81.

32. Vargas, D.M., De Bastiani, M.A., Zimmer, E.R. *et al.* Alzheimer's disease master regulators analysis: search for potential molecular targets and drug repositioning candidates. *Alz Res Therapy* **10,** 59 (2018).

33. Aksenov, M. Y., Tucker, H. M., Nair, P., Aksenova, M. V., Butterfield, D. A., Estus, S., & Markesbery, W. R. (1999). The expression of several mitochondrial and nuclear genes encoding the subunits of electron transport chain enzyme complexes, cytochrome c oxidase, and NADH dehydrogenase, in different brain regions in Alzheimer's disease. *Neurochemical research*, *24*(6), 767–774.

34. Stygelbout, V., Leroy, K., Pouillon, V., Ando, K., D'Amico, E., Jia, Y., Luo, H. R., Duyckaerts, C., Erneux, C., Schurmans, S., & Brion, J. P. (2014). Inositol trisphosphate 3-kinase B is increased in human Alzheimer brain and exacerbates mouse Alzheimer pathology. *Brain: a journal of neurology*, *137*(Pt 2), 537–552.

35. Wayne Chadwick, Randall Brenneman, Bronwen Martin, Stuart Maudsley, "Complex and Multidimensional Lipid Raft Alterations in a Murine Model of Alzheimer's Disease", *International Journal of Alzheimer's Disease*, vol. 2010, Article ID 604792, 56 pages, 2010. https://doi.org/10.4061/2010/604792

36. Saura, C. A., Parra-Damas, A., & Enriquez-Barreto, L. (2015). Gene expression parallels synaptic excitability and plasticity changes in Alzheimer's disease. *Frontiers in cellular neuroscience*, *9*, 318.

37. https://www.sciencedirect.com/science/article/pii/S0969996116301656#bb0950

38. Garcia-Esparcia, P., Sideris-Lampretsas, G., Hernandez-Ortega, K., Grau-Rivera, O., Sklaviadis, T., Gelpi, E., & Ferrer, I. (2017). Altered mechanisms of protein synthesis in frontal cortex in Alzheimer disease and a mouse model. *American journal of neurodegenerative disease*, *6*(2), 15–25.

39. https://thebiogrid.org/interaction/740378/apoe-ankh.html

40. Xiao, M. F., Xu, D., Craig, M. T., Pelkey, K. A., Chien, C. C., Shi, Y., Zhang, J., Resnick, S., Pletnikova, O., Salmon, D., Brewer, J., Edland, S., Wegiel, J., Tycko, B., Savonenko, A., Reeves, R. H., Troncoso, J. C., McBain, C. J., Galasko, D., & Worley, P. F. (2017). NPTX2 and cognitive dysfunction in Alzheimer's Disease. *eLife*, *6*, e23798.

41. Ramanan, V. K., Kim, S., Holohan, K., Shen, L., Nho, K., Risacher, S. L., Foroud, T. M., Mukherjee, S., Crane, P. K., Aisen, P. S., Petersen, R. C., Weiner, M. W., Saykin, A. J., & Alzheimer's Disease Neuroimaging Initiative (ADNI) (2012). Genome-wide pathway analysis of memory impairment in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort implicates gene candidates, canonical pathways, and networks. *Brain imaging and behavior*, *6*(4), 634–648.

42. Uhrig, M., Brechlin, P., Jahn, O. *et al.* Upregulation of CRABP1 in human neuroblastoma cells overproducing the Alzheimer-typical Aβ42reduces their differentiation potential. *BMC Med* **6,** 38 (2008).

43. Campanella, C., Pace, A., Caruso Bavisotto, C., Marzullo, P., Marino Gammazza, A., Buscemi, S., & Palumbo Piccionello, A. (2018). Heat Shock Proteins in Alzheimer's Disease: Role and Targeting. *International journal of molecular sciences*, *19*(9), 2603.

44. Stygelbout, V., Leroy, K., Pouillon, V., Ando, K., D'Amico, E., Jia, Y., Luo, H. R., Duyckaerts, C., Erneux, C., Schurmans, S., & Brion, J. P. (2014). Inositol trisphosphate 3-kinase B is increased in human Alzheimer brain and exacerbates mouse Alzheimer pathology. *Brain : a journal of neurology*, *137*(Pt 2), 537–552.

45. Lei, Z., Brizzee, C., & Johnson, G. V. (2015). BAG3 facilitates the clearance of endogenous tau in primary neurons. *Neurobiology of aging*, *36*(1), 241–248.

46. Bass, B., Upson, S., Roy, K., Montgomery, E. L., Jalonen, T. O., & Murray, I. V. (2015). Glycogen and amyloid-beta: key players in the shift from neuronal hyperactivity to hypoactivity observed in Alzheimer's disease?. *Neural regeneration research*, *10*(7), 1023–1025.

47. Wu, X., & Reddy, D. S. (2012). Integrins as receptor targets for neurological disorders. *Pharmacology & therapeutics*, *134*(1), 68–81.

48. Venkatasubramaniam, A., Drude, A., & Good, T. (2014). Role of N-terminal residues in Aβ interactions with integrin receptor and cell surface. *Biochimica et biophysica acta*, *1838*(10), 2568–2577.

49. Lechner, T., Adlassnig, C., Humpel, C., Kaufmann, W. A., Maier, H., Reinstadler-Kramer, K., Hinterhölzl, J., Mahata, S. K., Jellinger, K. A., & Marksteiner, J. (2004). Chromogranin peptides in Alzheimer's disease. *Experimental gerontology*, *39*(1), 101–113.

50. Xiao, M. F., Xu, D., Craig, M. T., Pelkey, K. A., Chien, C. C., Shi, Y., Zhang, J., Resnick, S., Pletnikova, O., Salmon, D., Brewer, J., Edland, S., Wegiel, J., Tycko, B., Savonenko, A., Reeves, R. H., Troncoso, J. C., McBain, C. J., Galasko, D., & Worley, P. F. (2017). NPTX2 and cognitive dysfunction in Alzheimer's Disease. *eLife*, *6*, e23798.

51. Rahman, M. R., Islam, T., Shahjaman, M., Zaman, T., Faruquee, H. M., Jamal, M., Huq, F., Quinn, J., & Moni, M. A. (2019). Discovering Biomarkers and Pathways Shared by Alzheimer's Disease and Ischemic Stroke to Identify Novel Therapeutic Targets. *Medicina (Kaunas, Lithuania)*, *55*(5), 191.

52. Zorzetto, M., Datturi, F., Divizia, L., Pistono, C., Campo, I., De Silvestri, A., Cuccia, M., & Ricevuti, G. (2017). Complement C4A and C4B Gene Copy Number Study in Alzheimer's Disease Patients. *Current Alzheimer research*, *14*(3), 303–308.

53. Iqbal, K., Liu, F., Gong, C. X., & Grundke-Iqbal, I. (2010). Tau in Alzheimer disease and related tauopathies. *Current Alzheimer research*, *7*(8), 656–664.

54. Zhu, D., & Van Meir, E. G. (2016). BAI1: from cancer to neurological disease. *Oncotarget*, *7*(14), 17288–17289.

55. Beckmann, N.D., Lin, WJ., Wang, M. *et al.* Multiscale causal networks identify VGF as a key regulator of Alzheimer's disease. *Nat Commun* **11,** 3942 (2020).